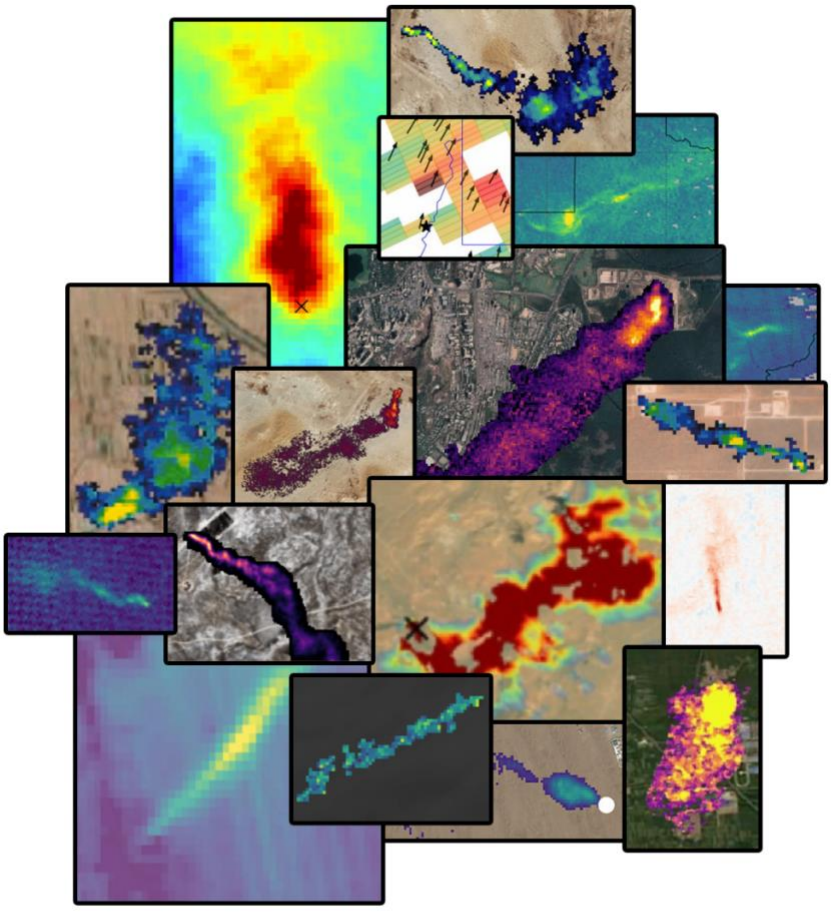
Algorithm Intercomparison Report (AIR)

WP320 - Deliverable D3.2



Authors

Lead authors:

- Michael Buchwitz: Document lead and FM (Flux Mapper) sections lead
- Javier Roger: HI (Hyperspectral Imager) sections lead
- Julian Akani-Guéry: BI (Band Imager) sections lead

All authors:

Institute of Environmental Physics, University of Bremen, Germany (IUP-UB)

- Michael Buchwitz (WP 310 lead, AIP document lead, FM lead)
- Oliver Schneising-Weigel
- Michael Hilker
- Stefan Noël
- Maximilian Reuter
- Jonas Hachmeister
- Hartmut Boesch

SRON Netherlands Institute for Space Research

- Matthieu Dogniaux
- Xin Zhang
- Joannes D. Maasakkers
- Ilse Aben

Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, Spain (UPV)

- Javier Roger (HI lead and UPV contribution to HI)
- Javier Gorroño (UPV contribution to BI)
- Luis Guanter (UPV contribution to HI and BI)

Kayrros, Paris, France

- Julian Akani-Guéry (BI lead)
- Quentin Peyle
- Maxence Nevoret

Change log

Version	Date	Status	Authors	Reason for change
Draft 1	11-Sept-2025	Submitted to team	(*)	New document
0.9	1-Oct-2025	Submitted to team for final check / inputs	(*)	Team inputs included
0.9.1	13-Oct-2025	Submitted to I. Aben (IA) for review	(*)	Team inputs included
0.9.2	10-Dec-2025	Submitted	(*)	Team inputs included
1.0	18-Feb-2026	Accepted	-	Version number change after acceptance

(*) see Authors list

Approved by

Date: 24-02-2026 Signature: 	Date: Signature:
Dr. Ilse Aben Science Lead, SRON	Dr. Simon Pinnock ESA Technical Officer

Table of Contents

1 Executive summary

Methane emission estimation results obtained from satellite data using different algorithms have been intercompared. This activity has been carried out in the framework of the ESA MEDUSA (Methane Emissions Detection Using Satellites Assessment) project using three different types of satellite instruments: (i) Flux Mappers (FM), which are city-scale resolution sensors (here: TROPOMI on the Sentinel-5 Precursor satellite), (ii) the Hyperspectral Imagers (HI) PRISMA, EnMAP, and EMIT, and (iii) the Band Imagers (BI) Sentinel-2A/B (S2A and S2B), Landsat-8/9 (L8 and L9), and GOES satellites.

For FM. i.e., TROPOMI, results from 3 algorithms have been compared: the Integrated Mass Enhancement (IME) algorithm of SRON, the modelling-based algorithm of Kayrros, and the Cross-Sectional-Flux (CSF) algorithm of IUP, University of Bremen (IUP-UB). Year 2021 data have been compared for selected target regions and globally. For the target regions, the CSF algorithm produced the largest number of emission estimates as it was applied to all suitable data instead of only to detected methane plumes like the other algorithms, followed by SRON and Kayrros. For the global comparison, good to reasonable agreement has been found for the common cases obtained by SRON and Kayrros with good correlation (R = 0.62) and low mean bias, but significant scatter of the relative differences (standard deviation: 53%, with 77% of plumes agreeing within uncertainty range) mainly driven by quantification method differences. These two algorithms include an automated plume detection step that also estimates the location of its emission source. SRON detected significantly more plumes (3233) compared to Kayrros (1060). The IUP-UB algorithm, where the location of interest is input, has only been applied to the selected targets. For several targets the IUP-UB estimates obtained with the CSF v1.6 algorithm were significantly higher compared to the SRON estimates. To investigate this, several changes have been implemented in CSF v1.7, e.g., a modified background correction. These modifications resulted in much improved agreement with SRON at the selected target locations.

For HI, i.e., PRISMA, EnMAP, and EMIT, the comparison has been carried out by addressing several aspects. Concerning “Plume identification and attribution to source” it has been found that the results from the plume delineation may differ significantly (though this should not affect quantification as long as the used calibrations are consistent). From task “Assessment of methane concentration retrievals used for detection” good consistency has been found, with the broad-wavelength combo matched filter showing improved performance for plume detection. Concerning “General methane emission quantification assessment” it has been found that the selection of the used wind speed values has a large impact on the quantification results. Specific aspects have been further investigated in the related task “Detailed methane emission quantification assessment”. It has been identified that an important fraction of the differences can be explained by differences in the methane enhancement maps, which mostly arise from the selection of the unit methane

absorption spectra (K spectra), and the specific criteria followed in the implementation of the quantification method. Several future research activities have been identified as not all questions could be answered. Where possible these activities will be carried out as part of other (continued) work packages or as part of other activities, e.g., in the context of planned peer-reviewed publications.

For BI, comparisons of results obtained with the Kayrros and UPV algorithms as applied to Sentinel-2 and Landsat-8/9 radiances have been carried out for extended time periods over target regions and for single scenes. The continuous monitoring intercomparison focused on two target regions. One is located in the Bugdayly gas field in Turkmenistan and the second target is a compressor station in the Permian Basin in the USA. The intercomparison focused on three aspects: (i) detection performance, (ii) methane emission quantification and (iii) analysis of missed plume detections. The target region intercomparison results can be summarized as follows: Concerning detection the following results have been obtained: For Sentinel-2, Kayrros' automated detection approach saves 85% of time, but at the cost of a 20% performance decrease compared to the manual method from UPV. The results are relatively similar for Landsat-8/9: 82% more time efficiency but with a more substantial performance trade-off (50% recall reduction). Concerning methane emission quantification the following has been found: Systematic differences of 76% between UPV and Kayrros Sentinel-2 plume quantifications, and 81% for Landsat-8/9. Both providers employ the IME methodology for flow rate quantification. The discrepancy originates from the integrated mass enhancement divided by characteristic length (IME/L). Concerning analysis of missed plume detections the following has been obtained: For Sentinel-2, no clear pattern emerges to explain plumes missed by either provider. For Landsat-8/9, Kayrros' neural network appears to regularly miss lower-intensity plumes (characterized by lower IME/L values). The single observations intercomparison results can be summarized as follows: Substantial discrepancies were observed between Landsat-8/9 and Sentinel-2 L2 products, primarily attributed to challenging observational conditions in the Iraq and Venezuela sites. These conditions introduced artifacts from gas flares, cloud contamination, and urban infrastructure, highlighting the critical importance of robust preprocessing methodologies and artifact removal procedures. The GOES intercomparison examines two case studies, an emission in Mexico and an emission in Idaho. For these events, the relative difference between UPV and Kayrros emission estimates is approximately 40%, with UPV estimates always higher than Kayrros, consistent with the magnitude observed for Sentinel-2 and Landsat-8/9 at two other sites (located in Turkmenistan and the Permian Basin in the USA). The analysis shows that the uncertainty intervals for the GOES quantification results do not overlap. This result stems from the significantly lower relative uncertainties achieved by both providers: UPV reported 5% and 6%, while Kayrros reported 3% and 6%. These low relative uncertainty values are due to the flow rate calculation methodology for geostationary satellites, which is performed

without relying on any external wind data. In general, it has been found that retrieval-driven biases dominate across all platforms and that mask effects are typically minor.

2 Detailed summary

This Algorithm Intercomparison Report (AIR) document provides an overview of the intercomparisons which have been carried out according to the MEDUSA Algorithm Intercomparison Plan (AIP) (Buchwitz et al., 2024). In addition to the algorithm intercomparison results it provides an overview of the considered satellite sensors and used plume detection and emission quantification algorithms.

The term “algorithm intercomparison” as used in this document primarily refers to comparisons of outputs as generated by the different emission detection and quantification algorithms as applied to different satellite input data sets. These outputs are primarily the derived emission estimates and their uncertainty but also additional data such as the underlying atmospheric methane fields, wind speed and direction, etc. The atmospheric methane fields are either existing input data sets or are generated specifically for the purpose of this project. The “emission algorithms” are also referred to as “Level 4” (L4) algorithms in this document and the resulting emission data products as L4 products. Input data for the L4 algorithms are Level 2 (L2) data products (atmospheric methane) or Level 1 (L1) products (radiances). L2 products are derived from L1 products. Based on the intercomparison plan the corresponding L4 products (including corresponding inputs and additional outputs) have been generated for specified periods and regions (for specific target regions and/or globally).

The algorithm intercomparison as described in this document is only one part of the MEDUSA project. It covers only a sub-set of the satellite sensors and emission estimation algorithms used in MEDUSA, namely those where more than one methane emission inversion algorithm is available (within this project) to obtain hotspot emission information for a given satellite sensor. In a parallel task, called “Systematic product intercomparison and validation” (WP200), additional sensors and algorithms have been used.

The algorithms and their intercomparisons as reported in this document are grouped into three groups depending on satellite instrument type. These three instrument types are:

- Flux Mapper (FM): These are moderate spatial resolution (several kilometer) sensors. In this task this is TROPOMI on Sentinel-5 Precursor (S5P).
- Hyperspectral Imager (HI): PRISMA, EnMAP and EMIT/ISS.
- Band Imager (BI): Sentinel-2 (S2), Landsat-8 (L8), Landsat-9 (L9) and GOES satellites.

Methods and intercomparison results are reported per instrument type, i.e., separately for FM, HI and BI.

This document consists of a main document and 3 Annexes. The main document provides an overview and summarizes the results referring for details on FM methods and results to Annex A, for HI to Annex B, and for BI methods and results to Annex C.

The intercomparison results for FM (i.e., TROPOMI) comprise an intercomparison for ten target locations and a global intercomparison that can be summarized as follows (for details see also Annex A):

2.1 Summary of FM results

Detailed comparisons of methane emission estimates (Level 4 products) at 10 target locations have been presented in this section as obtained from TROPOMI satellite XCH_4 retrievals (Level 2 products) using the 3 different Level 2 to Level 4 algorithms as developed and used by SRON, Kayrros, and IUP-UB. The comparison shows that, as expected, the IUP-UB algorithm typically provides the largest number of methane emission estimates for a given target location. This is because the IUP-UB Cross-Sectional-Flux (CSF) algorithm is not based on emission plume detection in contrast to the other 2 algorithms. Instead, only a small set of selection criteria is used by IUP-UB in order to decide if the satellite data can be used for emission estimation or not: Data coverage must be dense enough (e.g., > 90% of the area used for the CSF analysis needs to be covered by “good” (e.g., cloud free) data independent of the spatial pattern of the methane concentration field) and wind speed should not be too low (> 1 m/s). Consequently, estimated emissions using IUP-UB’s CSF algorithm may also be zero or even negative in case no methane plume is present and because XCH_4 retrievals are noisy. The smallest number of emission estimates are provided by the Kayrros algorithm due to a conservative detection method and their quantification approach, which is based on detailed modelling, which is challenging due to the required good knowledge of meteorological information (e.g., wind field), issues related to turbulence, required match between modelled and observed plume, etc. The quantification method can therefore not be applied to all detected plumes. The number of estimates provided by SRON is in between the number of estimates of Kayrros and IUP-UB.

The IUP-UB CSF algorithm as used for MEDUSA is new and still under development. Most of the results generated by IUP-UB for MEDUSA have been produced with version 1.6 of this algorithm. CSF v1.6 is also the algorithm used for emission estimates from the Hyperspectral Imagers (HI) EnMAP, PRISMA and EMIT (see Annex B to document AIR). Results reported in the following refer to comparisons with v1.6 except if stated otherwise.

Primarily due to the low number of Kayrros estimates and the fact that often only one algorithm provided an estimate for a given target location and satellite overpass, the number of collocations is small. At 3 of the 10 target locations no collocations have been found. For the 7 other locations where collocations exist, on average good to reasonable agreement between the IUP-UB and the SRON estimates has been found for 3 of these targets but for the 4 other targets, the IUP-UB estimates were significantly

higher (on average by more than a factor of 2). In contrast, the agreement between Kayrros and SRON was found to be much better with agreement typically within ~10%. But this finding is based on only a very small number of collocations and the global comparisons results reported separately are more meaningful.

The small number of targets and small number of collocations per target raises the question of how representative the target region comparisons are. Additional comparisons for several other locations have been conducted independently by B. Dils (BIRA-IASB) in the context of MEDUSA work package (WP) 220. These comparisons show on average very good agreement between the emission estimates as obtained by IUP-UB with CSF v1.6, SRON and Kayrros. There, the main findings based on correlation (scatter) plots are: slope $\beta = 1.01$ ($N = 26$, $R^2 = 0.74$) for SRON vs IUP-UB, $\beta = 0.92$ ($N = 22$, $R^2 = 0.69$) for Kayrros vs IUP-UB and $\beta = 1.05$ ($N = 22$, $R^2 = 0.55$) for SRON vs Kayrros. This indicates that all three algorithms result in very similar emission estimates on average (as all slopes are close to 1) although (as indicated by the moderate R^2 values) the individual emission estimates may vary significantly (by often several 10%). This indicates that the level of agreement or disagreement of the results obtained with the different algorithms depends on which subset is compared.

Several assessments have nevertheless been carried out to explain the relative overestimation of the IUP-UB CSF v1.6 emission estimates compared to SRON at some of the target locations. To investigate the potential impact of the use of different TROPOMI Level 2 (L2) input data, the IUP-UB CSF algorithm has also been applied to the operational L2 data (instead of using the WFMD v1.8 L2 product, which are the default input data for IUP-UB). The analysis revealed that the different Level 2 input data are not the driver for the emission differences. Also, the wind speed used by IUP-UB is unlikely the reason for a factor of 2 or larger differences. IUP-UB uses for TROPOMI the Planetary Boundary Layer (PBL) average of the wind speed, which is in line with Varon et al., 2019, who found that the PBL average wind speed was appropriate for TROPOMI CSF-based emission estimates for a strongly emitting location in Turkmenistan. Additional sensitivity studies have resulted in a new version 1.7 of the IUP-UB CSF algorithm. The main differences of v1.7 compared to v1.6 are a modified background correction and a waiver of the 3-hour plume length limit used for v1.6. The waiver of the 3-hour limit typically results in a reduction of the emission estimates as typically the computed fluxes through the cross-sections are lower for cross-sections further away. In this context it is also important to know how the reported emissions need to be interpreted. For v1.6, with the 3-hour limit implemented, the reported emission referred to the average emission in the period 3 hours before the satellite overpass to the time of the satellite overpass. For v1.7 this averaging time is not constant anymore but depends on wind speed. For an emission source emitting at a constant rate both approaches lead (ideally) to identical results but this is not true for sources with time dependent emission. It is shown that v1.7 results in much better agreement with the SRON estimates at the selected target locations indicating that the addressed aspects may explain at least part of the observed differences. So far only a limited amount of data has been processed and analyzed with v1.7 and further studies are needed to find out if v1.7 is better than v1.6 also for cases not yet investigated.

IUP-UB is working on this aspect, e.g., in the context of preparing for a peer reviewed publication.

Due to the limited number of collocations with the Kayrros estimates at the selected 10 target regions, the target region comparison focused on comparisons between the IUP-UB and SRON algorithms. For a more meaningful comparison of the SRON and the Kayrros algorithms, please see the following summarizing the global comparisons.

The Global comparison results can be summarized as follows:

Global datasets of methane emission plumes detected in TROPOMI data have been reported by Kayrros and SRON for the year 2021. Both datasets have detections covering the main methane emitting regions (e.g., the Permian basin, West coast of Turkmenistan, Shanxi coal mines, etc.), with SRON also covering additional locations (e.g. Casablanca, Istanbul, and coal mines in Poland) consistent with a three-times higher number of detections (3233 SRON versus 1060 Kayrros). Plumes from both datasets have been matched based on the comparison of their reported plume masks. The Kayrros – SRON matching plumes (N=643, with 306 that include emission quantifications in both) agree on average (small mean difference) and show a good correlation (r = 0.62). This result is quite remarkable considering that Kayrros and SRON quantification methods are strictly different. However, the scatter of Kayrros – SRON differences is quite significant, with a 53% relative difference standard deviation. Emission estimates for 77% of the 306 plumes agree within uncertainty range. Most of the plume-to-plume emission difference scatter is shown to be explained by the difference between Kayrros and SRON emission quantification methods: using SRON's quantification method on Kayrros detected plumes reduces the difference variance by 83%. This comparison of Kayrros – SRON matching plumes also brought up useful improvement points for both algorithms: SRON's effective wind calibration are slightly inconsistent between 10-m and PBL-averaged wind speeds; Kayrros' handling of missing TROPOMI pixels appears to worsen comparison results. Examining the 386 plumes detected by Kayrros but not reported by SRON, we assess that filtering (39% due to TROPOMI pixel filters and 35% due to human labelling of detections) employed by SRON can explain the large majority of these non-detected or non-reported plumes. Conversely, of the 2568 SRON-only detections, 70% correspond to short and low-intensity plumes, which Kayrros' anomaly-based computer vision algorithm tends to miss due to its reliance on smoothed input data. Additional differences (15%) stem from borderline cases, where variations in pipeline design and input data can cause detections to fall into different filtering categories across teams.

2.2 Summary of HI results

The intercomparison results for HI can be summarized as follows (for details see also Annex B):

For HI, i.e., PRISMA, EnMAP, and EMIT/ISS, the comparison has been carried out by addressing four different aspects: (i) Plume identification and attribution to source, (ii) Assessment of methane concentration retrievals used for detection, (iii) General methane emission quantification assessment, and (iv) Detailed methane emission quantification assessment. For each of these four aspects, detailed results have been produced and discussed, and potential further research aspects have been identified. Concerning “Plume identification and attribution to source”, it has been found that the results from the plume delineation may differ significantly depending on the approach and calibration used. From the task “Assessment of methane concentration retrievals used for detection” good consistency has been found. It has also been found that the spectral window used for retrieval is relevant for the used Matched Filter (MF) methods, with the “Combo-MF” using a wider wavelength range showing better performance for plume detection. Concerning “General methane emission quantification assessment”, it has been found that the selection of the used wind speed values has a large impact on the quantification results as available meteo products differed by up to 2 m/s for the same plume case. Specific additional aspects have been further investigated in the related task “Detailed methane emission quantification assessment”. It has been identified that an important fraction of differences for a few plumes studied in detail can be explained by differences in the methane enhancement maps, which mostly arise from the selection of the unit methane absorption spectra (K spectra), and the specific criteria followed for the implementation of the quantification method. Several aspects for future research have been identified as not all questions could be answered, for example, why the IUP-UB CSF-based emission estimate for an EMIT scene in Kansas is significantly lower compared to the estimates obtained via the other algorithms.

2.3 Summary of BI results

The BI intercomparison consists of a continuous monitoring intercomparison as well as an intercomparison of single scenes. Results can be summarized as follows (for details see also Annex C):

Continuous monitoring intercomparison results:

The continuous monitoring intercomparison focused on two target regions. One is in the Bugdayly gas field in Turkmenistan, where two sub-regions have been identified for detailed study referred to as TKM1 and TKM2. The second target is a compressor station in the Permian Basin located in the USA.

The intercomparison focused on three aspects: (i) Detection performance, (ii) Methane emission quantification, and (iii) Analysis of missed plume detections. The findings can be summarized as follows:

Detection performance: For Sentinel-2, Kayrros' automated detection approach saves 85% of time, but at the cost of a 20% performance decrease compared to the UPV manual method. The observation is relatively similar for Landsat-8/9: 82% more time efficiency but with a more substantial performance trade-off (50% recall reduction).

Methane emission quantification: We observed systematic differences of 76% between UPV and Kayrros Sentinel-2 plume quantifications, and 81% for Landsat-8/9. Both providers employ the IME methodology for flow rate quantification. The discrepancy originates from either the effective wind (Ueff) or the integrated mass enhancement divided by a characteristic length (IME/L):

- *Ueff:* Both providers use the same wind source and effective wind formula. The only difference is that Kayrros selects the nearest wind point from the past, while UPV applies temporal interpolation. As anticipated, this produces minimal differences (~5%), except for two cases showing 14% relative error.
- *IME/L:* This parameter accounts for most of the systematic bias in both Sentinel-2 and Landsat-8/9 quantifications. To investigate further and determine whether the mask or L2 product drives this difference, we selected individual scenes for detailed analysis. Ground-truth data from sources like Stanford controlled release campaigns is needed to assess which approach is best.

Analysis of missed plume detections: For Sentinel-2, no clear pattern emerges to explain plumes missed by either provider. For Landsat-8/9, Kayrros' neural network appears to regularly miss lower-intensity plumes (characterized by lower IME/L values).

Single observations intercomparison results:

Three scenes have been identified for detailed intercomparison for Sentinel-2 and Landsat corresponding to oil and gas fields in Iraq and Venezuela. The Iraq scene is a desert scene with a complex heterogeneous background, e.g., due to many flaring stacks. The chosen Venezuela site is characterized by dense vegetation that changes seasonally, frequent cloud coverage and a lot of flaring. These conditions introduced artifacts from gas flares, cloud contamination, and urban infrastructure, highlighting the critical importance of robust preprocessing methodologies and artifact removal procedures. Future work should focus on identifying whether observed differences originate from background removal techniques or methane retrieval algorithms applied to residual or reflectance ratio data. The GOES intercomparison examines two case studies, an emission in Mexico and an emission in Idaho. For these events, the relative difference between UPV and Kayrros estimates is approximately 40%, consistent with the magnitude observed for Sentinel-2 and Landsat-8/9. The analysis shows that the uncertainty intervals for the GOES quantification results do not overlap. This result stems from the significantly lower relative uncertainties achieved by both providers: UPV reported 5% and 6%, while Kayrros reported 3% and 6%. These low relative uncertainty values are due to the

flow rate calculation methodology for geostationary satellites, which is performed without relying on any external wind data. In general, it has been found that retrieval-driven biases dominate across all platforms, and that mask effects are typically minor.

3 Overview satellite sensors

This section provides a short overview of the satellite sensors relevant for the MEDUSA algorithm intercomparison. These are sensors where more than one methane emission inversion algorithm is used and evaluated within MEDUSA. This is a sub-set of the satellite sensors used within MEDUSA. **Table 3-1** presents an overview of the used satellite instruments. Details are provided in the following sub-sections.

Instrument Type	Instrument / Satellite	Launch date	Comments
FM	TROPOMI/S5P	Oct-2017	
HI	PRISMA	Mar-2019	
HI	EnMAP	Apr-2022	
HI	EMIT on ISS	July-2022	
BI	Landsat 8 (L8) Landsat 9 (L9)	Feb-2013 Sept-2021	
BI	Sentinel 2 (S2) series	2A: Jun-2015 2B: Mar-2017	
BI	GOES with ABI instrum. GOES-R (or GOES-16) GOES-S (or GOES-17) GOES-T (or GOES-18) GOES-U (or GOES-19)	 Nov-2016 Mar-2018 Mar-2022 Jun-2024	 Cent. America, East. US Pacific ocean Cent. America, West. US Atlantic ocean

Table 3-1: *Overview satellite instruments as used for the MEDUSA algorithm intercomparison. Note that only GOES-R is used in this part of the MEDUSA project.*

3.1 Flux Mapper (FM)

3.1.1 TROPOMI on Sentinel-5 Precursor (S5P)

TROPOMI is the only payload instrument of the ESA Sentinel-5 Precursor mission launched in October 2017 (Veefkind et al., 2012). It is a push-broom imaging spectrometer measuring radiances in eight spectral bands from the ultraviolet (UV) to the shortwave infrared (SWIR), allowing the retrieval of various atmospheric constituents including methane. TROPOMI has the unique ability to combine observations at high spatial resolution with global coverage on a daily basis due to its

large 2600 km swath consisting of individual measurements with a footprint size of 5.5 x 7 km^2 at nadir (7x7 km^2 before 6 August 2019) in the SWIR bands (7 and 8) relevant for methane retrieval.

Input for this study are Level 2 (L2) XCH_4 data products from TROPOMI/S5P, where XCH_4 is the column-averaged dry-air mole fraction of atmospheric methane. We have used the operational data product, which is based on the retrieval code developed by SRON (e.g., Lorente et al., 2021, 2023) and the scientific data product generated by University of Bremen using the WFMD (Weighting Function Modified DOAS) retrieval algorithm (e.g., Schneising et al., 2023).

3.2 Hyperspectral Imagers (HI)

3.2.1 PRISMA

PRISMA (PRecursore IperSpettrale della Missione Applicativa) is a hyperspectral imaging mission operated by the Italian Space Agency (ASI). Launched in March 2019, PRISMA carries a push-broom hyperspectral sensor capable of capturing images in 239 spectral bands from 400 nm to 2500 nm. The instrument provides detailed spectral information with a spectral resolution and sampling better than 12 nm. The spatial resolution is 30 m for the hyperspectral bands and 5 m for the panchromatic band. PRISMA is mainly operated by user requests for data acquisition. The high spectral resolution across a broad wavelength range allows for detailed analysis of material composition, while the panchromatic band enhances spatial details and aids in sharpening hyperspectral imagery. After registration data can be downloaded from the PRISMA website (https://prisma.asi.it).

3.2.2 EnMAP

EnMAP (Environmental Mapping and Analysis Program) is a German hyperspectral imaging mission managed by the German Aerospace Center (DLR). Launched in April 2022, EnMAP features a hyperspectral imager with a push-broom sensor that captures data in 242 spectral bands across the VNIR (420–1000 nm) and SWIR (900–2450 nm) ranges. The spectral resolution and sampling are approximately 6.5 nm in the VNIR range and around 10 nm in the SWIR range. Users can request EnMAP data acquisitions for specific areas of interest by data tasking. The high signal-to-noise ratio and continuous coverage from VNIR to SWIR enable precise detection of surface features and materials, facilitating comprehensive environmental monitoring and analysis. After registration data can be downloaded from the EnMAP website (https://www.enmap.org/data_access/).

3.2.3 EMIT

EMIT (Earth Surface Mineral Dust Source Investigation) is an imaging spectrometer mission developed by NASA's Jet Propulsion Laboratory (JPL). Launched in July 2022, EMIT initially focused on mapping the composition of mineral dust in arid regions to better understand its effects on climate. The instrument captures data in the VNIR to SWIR range (380–2500 nm) with hundreds of narrow spectral bands. The spectral resolution and sampling are approximately 7.5 nm, while the spatial resolution is 60 m. EMIT's revisit time depends on its position on the International Space Station (ISS) and the specific target areas but generally ranges from a few days to weeks. Data can be downloaded from the EMIT website (https://earth.jpl.nasa.gov/emit/data/data-products/).

3.3 Band Imager (BI)

3.3.1 Sentinel-2 (S2 series)

Sentinel-2 is a multispectral imaging mission operated by ESA with three satellites: Sentinel-2A launched in June 2015, Sentinel-2B launched in March 2017. Sentinel-C in June 2024. While Sentinel-2A was originally scheduled for decommissioning and replacement by Sentinel-2C, an exceptional one-year extension campaign began in March 2025. As our analysis focuses on imagery before June 2024, Sentinel-2C is excluded from this study. Each of them carries a push-broom instrument measuring 13 spectral bands: 4 VNIR bands at 10 m spatial resolution, 6 SWIR bands at 20 m, and 3 bands at 60m for atmospheric correction purposes and cloud screening. The combination of the two satellites (Sentinel-2A and Sentinel-2B in our case) provides global coverage with a 5-day revisit time.

3.3.2 Landsat-8 (L8) and Landsat-9 (L9)

Landsat 8 and 9 are part of NASA's Landsat program and were respectively launched in February 2013 and September 2021. They carry a push-broom imager measuring 9 spectral bands between 433 and 2300 nm: 5 VNIR and 3 SWIR bands at 30 m spatial resolution, and 1 panchromatic band at 15 m spatial resolution. The combination of the two satellites provides global coverage in an 8-day revisit time.

3.3.3 GOES

GOES 3rd generation series gathers four geostationary satellites developed by NOAA and NASA.

- GOES-R (or GOES-16) was launched in November 2016 and is located above Central America and the Eastern US.

- GOES-S (or GOES-17) was launched in March 2018 and is located above the Pacific Ocean.
- GOES-T (or GOES-18) was launched in March 2022 and is located above Central America and the Western US.
- GOES-U (or GOES-19) was launched in June 2024 and is located above the Atlantic Ocean.

These satellites carry among other instruments an Advanced Baseline Imager (ABI), which is a multispectral sensor. It measures 2 visible bands at 500 m spatial resolution at nadir, 3 short wave infrared (SWIR) bands at 1 km and 11 medium wave infrared (MWIR) bands at 2 km. Only GOES-R imagery is used in this study.
GOES-R routinely operates following two acquisition modes:

- The CONUS (Continental United States) mode corresponds to the acquisition of an image every 5 minutes on the CONUS sector defined by the US, and parts of Canada and Mexico.
- The Full Disk mode corresponds to the acquisition of an image every 15 minutes on the Earth disk visible from the satellite.

These satellites offer unique temporal resolution compared to Low Orbiting satellites.

4 Overview methane plume detection & emission rate estimation algorithms

This section provides a brief overview of methane plume detection and emission rate estimation algorithms as used and assessed within the MEDUSA algorithm intercomparison task.

In this document algorithms to obtain methane emission information are also referred to as Level 4 (L4) algorithms. These algorithms use as input lower-level data products such as Level 2 (atmospheric methane information from and for individual satellite footprints or ground pixel) or Level 1 (i.e., radiance) products.

For a general overview on methods and community-accepted practices for quantifying methane emissions based on plumes detected via spectroscopic remote sensing please see Worden et al., 2025, from which **Fig. 4-1** was taken, or Jacob et al., 2022.

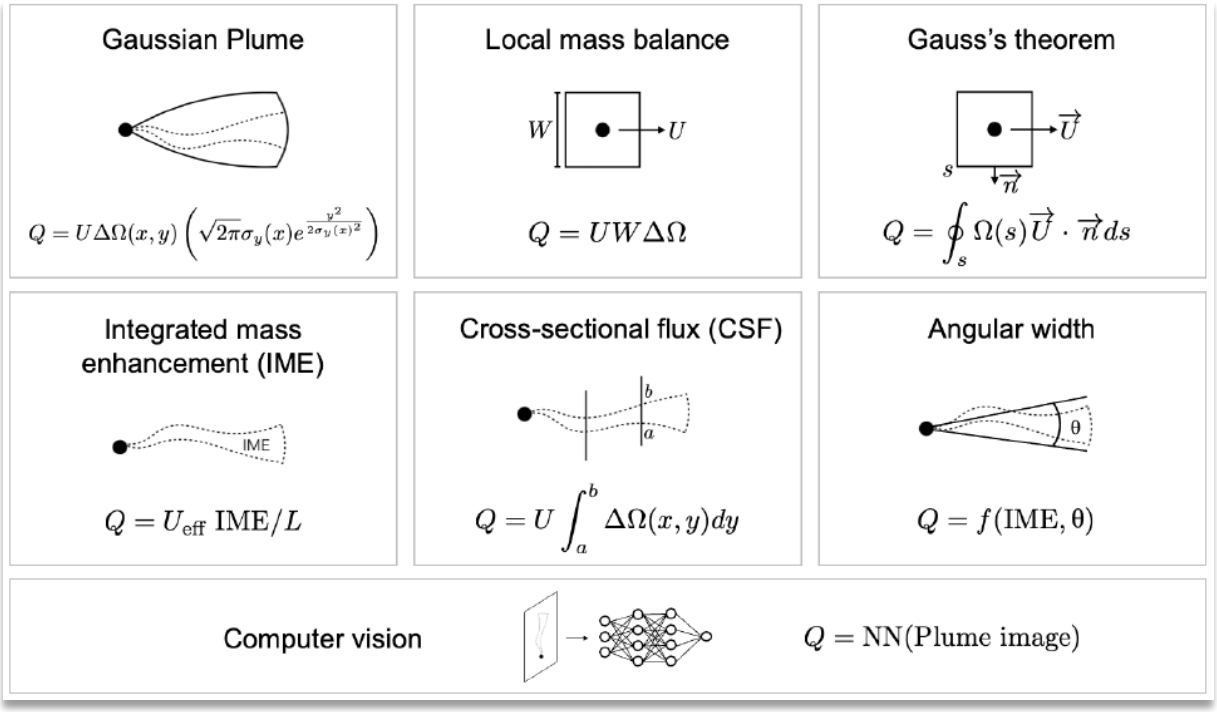


Figure 4-1: *Overview of methods used to obtain methane emission estimates (Q) from atmospheric methane vertical columns (Ω) or column enhancements (ΔΩ) and wind speed (U). Figure from Worden et al., 2025 (based on Jacob et al., 2022).*

Table 4-1 presents an overview of the algorithms used for instrument type FM, **Tab. 4-2** for HI and **Tab. 4-3** for instrument type BI.

Inst. Type	Inst./Sat.	Institute	Input product	Algorithm: Plume detection: Automatic? / Method	Algorithm: Emission quantification
FM	S5P	SRON	L2 operational	Yes / ML	IME (§)
FM	S5P	Kayrros	L2 operational	Yes / DL	Inversion
FM	S5P	IUP-UB	L2 WFMD	No / VI (#)	CSF

Table 4-1: *Overview of algorithms as used for instrument type FM. Abbreviations: WFMD: Weighting Function Modified DOAS algorithm; ML: Machine Learning; DL: Deep Learning; VI: Visual Inspection. IME: Integrated Mass Enhancement method; CSF: Cross-Sectional Flux method. (#) IUP-UB: Emission estimates for MEDUSA only for target regions (target location is input for current algorithms). (§) SRON: IME is baseline; other algorithms possible (CSF, atmospheric inversion).*

Inst. Type	Inst./Sat.	Institute	Input product	Algorithm: Plume detection: Automatic? / Method	Algorithm: Emission quantification
HI	PRISMA	SRON	L1 oper.	No / VI	IME
HI	PRISMA	UPV	L1 oper.	No / VI	IME
HI	PRISMA	Kayrros	L1 oper.	No / VI	IME
HI	PRISMA	IUP-UB	L1 oper.	No / VI	CSF
HI	EnMAP	SRON	L1 oper.	No / VI	IME
HI	EnMAP	UPV	L1 oper.	No / VI	IME
HI	EnMAP	Kayrros	L1 oper.	No / VI	IME
HI	EnMAP	IUP-UB	L1 oper.	No / VI	CSF
HI	EMIT	SRON	L1 oper.	No / VI	IME
HI	EMIT	UPV	L1 oper.	No / VI	IME
HI	EMIT	Kayrros	L1 oper.	Yes / DL	IME

Table 4-2: *Overview of algorithms as used for instrument type HI. Abbreviations: DL: Deep Learning; VI: Visual Inspection. IME: Integrated Mass Enhancement method; CSF: Cross-Sectional Flux method.*

Inst. Type	Inst./Sat.	Institute	Input product	Algorithm: Plume detection: Automatic? / Method	Algorithm: Emission quantification
BI	S2 A&B	Kayrros	L1 oper.	Yes / DL	IME
BI	S2 A&B	UPV	L1 oper.	No / VI	IME
BI	L8 & L9	Kayrros	L1 oper.	Yes / DL	IME
BI	L8 & L9	UPV	L1 oper.	No / VI	CSF
BI	GOES	Kayrros	L1 oper.	Yes / DL	IME
BI	GOES	UPV	L1 oper.	No / VI	IME

Table 4-3: *Overview of algorithms as used for instrument type BI. Abbreviations: DL: Deep Learning; VI: Visual Inspection. IME: Integrated Mass Enhancement method; CSF: Cross-Sectional Flux method.*

In this section only a short overview is presented. For details see (see **Sect. 9**):

- ANNEX A for FM algorithms
- ANNEX B for HI algorithms
- ANNEX C for BI algorithms

4.1 Algorithms for FM

Here only a short overview of the FM algorithms is presented. For details see ANNEX A (see **Sect. 9**).

4.1.1 SRON algorithm for TROPOMI/S5P

The base SRON TROPOMI methane plume estimation algorithm is based on the Integrated Mass Enhancement (IME) method, implemented to quantify plumes automatically detected in TROPOMI data through a 2-step machine learning procedure. All detections are manually verified by at least 2 persons. This scheme is extensively described in Schuit et al., 2023. The quantification method relies on the widely used Integrated Mass Enhancement (IME) method, first introduced by Frankenberg et al., 2016, for aircraft measurements, and first applied to satellite data by Varon et al., 2018. It relates the plume emission rate Q with the excess of methane contained in the plume mask denoted as IME. Plume length L is computed as square root of the area of the plume. Effective wind speed U_{eff} was calibrated via atmospheric transport simulations against the 10m wind speed (U_{10m}) as well as the wind speed averaged over the Planetary Boundary Layer thickness (U_{PBL}). An ensemble approach varying input data (such as the used wind product) is used to estimate a plume's emission rate and the associated uncertainty.

4.1.2 Kayrros algorithm for TROPOMI/S5P

Kayrros algorithm is based on the computation of a methane anomaly map, described in Lauvaux et al., 2022, applied to the TROPOMI/S5P operational L2 data V02.05.00. Continuous groups of positive pixels in the anomaly map are selected as plume candidates, and contiguous (but distinct) plumes are separated using watershed segmentation. The detections are systematically validated by two human expert labelers who filter out false positives. Each detection is quantified using a corresponding simulated plume, calculated with the Lagrangian particle dispersion model HYSPLIT (Stein et al., 2015). The emission rate is finally obtained by normalizing the flow rate of the simulated plume with the ratio of enhancement between the actual observation and the simulation. A flux rate uncertainty is estimated through a sensitivity analysis conducted on a batch of 200 plumes, evaluating the uncertainty arising from Sentinel-5P measurements, source location selection, background quantification, and meteorological data errors.

4.1.3 IUP-UB algorithm for TROPOMI/S5P

The IUP-UB methane emission (i.e., L4) algorithm is based on the Cross-Sectional-Flux (CSF) approach. Atmospheric methane (Level 2) input data from TROPOMI/S5P is the scientific XCH_4 data product version 1.8 as generated with the Weighting Function Modified Differential Absorption Spectroscopy (WFM-DOAS or WFMD) retrieval algorithm (Schneising et al., 2023). The IUP-UB CSF algorithm is applied to TROPOMI XCH_4 retrievals around pre-defined target locations. A plume detection algorithm is therefore not used but also not needed. For wind information meteorological information from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis product (Hersbach et al., 2018) is used. For TROPOMI the Planetary Boundary Layer (PBL) average wind speed and direction and its variability is used. The main result of this CSF algorithm is an estimate of the methane emission including 1-sigma uncertainty for given target locations and for each satellite overpass, where enough (quality filtered) data are available near the target location. Other filtering criteria are also used, e.g., the wind speed must be larger than a certain threshold (> 1 m/s). Emissions are estimated for several cross-sections (each corresponding to a different distance from the target) using all data in a box of size 100 km times 100 km at the target. The reported emission is the average of the individual estimates. Its corresponding (1-sigma) uncertainty is computed considering several contributions such as the standard deviation of the individual estimates, the number of independent cross-sections and wind speed variability. The emission product file is in CSV format with each line corresponding to a single satellite overpass containing at least the following information: target latitude and longitude, time, estimated emission and its (1-sigma) uncertainty (both in tons/hour) and the quality flag (0 = good). “Good” quality requires that the following conditions are met: (i) enough data near the target (coverage > 90%), (ii) wind speed not too low (> 1 m/s) and (iii) at least one independent cross-section.

4.2 Algorithms for HI

Here only a short overview of the HI algorithms is presented. For details see ANNEX B (see **Sect. 9**).

1.1.1. SRON algorithm for PRISMA, EnMAP, and EMIT

The SRON algorithm applies the matched filter using different wavelength windows to retrieve methane enhancements for plume detection and quantification. To determine the unit methane absorption spectrum (k), SRON employs a forward model (Gloudeman et al., 2008) and convolves the radiance with the imager's central wavelength and FWHM. To mitigate the background noise in urban areas for detections, we perform the matched filter over the entire near-infrared window (1300-2500 nm). Then, SRON applies a Chambolle total variation (TV) denoising filter

(Chambolle. 2004) to obtain a denoised methane enhancement field. Plume detection is carried out by automasking the denoised methane enhancement field and visual inspection. Finally, the retrieval is run again using the 2110–2450 nm band while removing plume pixels from the background estimation to make sure the sparsity assumption of the matched filter is not broken and using a lognormal matched filter (Pei et al., 2023) if the emission rate is higher than 10 t h^{-1}. The algorithm uses the Integrated Mass Enhancement (IME) method to quantify the emission rate. SRON performs instrument-specific calibrations for the effective wind speed (U_{eff}) based on large-eddy simulations.

4.2.1 Kayrros algorithm for PRISMA, EnMAP, and EMIT

The Kayrros algorithm is based on the matched filter formulation described in Guanter et al., 2021, applied to the L1B hyperspectral product. The matched filter is performed on a per-column basis on the spectral ranges 1550 nm - 1800 nm and 2100 nm – 2480 nm where methane respectively has weak and strong absorption features. For every scene, a high-resolution reference spectrum is generated by integrating the Beer-Lambert law on an atmosphere model of homogeneous layers. The unit absorption spectrum is eventually computed as the Jacobian of the reference absorption spectrum for an additional ppm of methane. The plume detection and delineation are automatically performed by a Deep Learning model (Groshenry et al., 2022). Every detection is quantified using the Integrated Mass Enhancement (IME) method (Varon et al., 2018). The ERA5 10m wind speed reanalysis data (Hersbach et al., 2018) is used, assuming a default 0.5 m/s uncertainty on each component of the wind vector. The wind speed is converted into effective wind speed according to the formula from Guanter et al., 2021.

4.2.2 UPV algorithm for PRISMA, EnMAP, and EMIT

UPV filters out the radiance values related to clouds and their shadows. They do the same with water if potential plumes are in land, and with the land pixels if they are in water. We identify these pixels by a manually selected radiance threshold set by visual inspection. The methodology used to retrieve methane concentration enhancement maps is the matched-filter (MF) method and related flavors. For plume detection, UPV uses Combo-MF as in Roger et al., 2024, and the simple matched-filter as in Guanter et al., 2021, for emissions coming from area sources. On the other hand, for quantification, they use an in-house lognormal MF (LMF) based on Pei et al., 2023, which considers two LMF retrievals that use different unit methane absorption spectra (K). K is deduced using a lookup table that relates methane transmittance spectra obtained using the MODTRAN radiative transfer code to different values of methane column mixing ratio, while accounting for the specific angular configuration at measurement time. Regarding the spectral window of application of the applied methodology, they use 2100 – 2400 nm as in Guanter et al., 2021, for every mentioned method, except for the Combo-MF, which uses a 1000 – 2500 nm spectral range.

Plume detection is made by visual inspection in those retrievals used for detection. The quantification is made using the Integrated Mass Enhancement (IME) method. The flux rate emission uncertainty is obtained by the square root of the quadratic sum of the errors related to the IME and the Ueff values.

4.2.3 IUP-UB algorithm for PRISMA and EnMAP

The IUP-UB methane emission estimation algorithm is based on the Cross-Sectional-Flux (CSF) approach. The algorithm is essentially identical to the IUP-UB algorithm described above for FM, i.e., TROPOMI (see **Sect. 4.1.3**). The only differences are:

- The box size is 500 m times 500 m (compared to 100 x 100 km^2 as used for TROPOMI)
- The 10 m wind is used (instead of the PBL average)

The HI methane image input data have been retrieved from the satellite Level 1B files using the Matched Filter (MF) method (see Bösch and Hilker, 2024).

4.3 Algorithms for BI

Here only a short overview of the BI algorithms is presented. For details see ANNEX C (see **Sect. 9**).

4.3.1 Kayrros algorithm for Sentinel-2, Landsat 8/9, GOES

These Band Imagers (BIs) have two band measurements in the SWIR domain: a band B_{strong} where methane strongly absorbs (around 2200 nm) and a band B_{weak} where methane absorbs much less (around 1600 nm). Kayrros algorithm leverages correlation between these two bands for other absorbing materials to enhance methane by computing the band ratio $\frac{B_{weak}}{B_{strong}}$. To further enhance methane, images from previous observations are used to estimate and remove the background. Only previous observations are used since the algorithm is running on a live basis as soon as the satellite image is available. The images obtained after background removal are called methane residuals. The detection and plume delineation are performed in an automated way by a deep learning model on the methane residuals. Emissions are quantified using the Integrated Mass Enhancement (IME) method. For specific aspects related to GOES see Groshenry et al., 2024, and ANNEX C.

4.3.2 UPV algorithm for Sentinel-2 A&B, Landsat 8/9, GOES

The UPV BI algorithm involves a ratio of bands together with temporal normalisation. The methane enhancement transmittance is approximated as the ratio between the radiance at a spectral channel affected by methane absorption (this is B12 for Sentinel

2 and B7 for Landsat) and a methane-free reference band (this is B11 for Sentinel 2 and B6 for Landsat). The next step seeks to relate the transmittance image to a methane concentration enhancement ΔXCH_4. This includes a radiative transfer look-up-table (LUT) scheme. The methane enhancement maps $\Delta XCH4$ are further converted into methane flux rates Q by the means of the Integrated Mass Enhancement (IME) method (Frankenberg et al., 2016; Varon et al., 2018).

5 Intercomparison results for Flux Mapper (FM)

This section summarizes the intercomparison results for the FM, i.e., for the algorithms used to obtain methane emission information from TROPOMI/S5P. All details are reported in ANNEX A (see **Sect. 9**).

5.1 Intercomparison results for target regions

5.1.1 Purpose and approach

To compare in detail the methane emission estimates obtained from the different algorithms, as applied to TROPOMI/S5P XCH_4 retrievals, 10 different target regions have been defined covering several interesting and important methane sources such as landfills, oil and gas fields, and coal mines. Results from 3 different algorithms have been compared: the IME algorithm of SRON, the modelling-based algorithm of Kayrros and the CSF algorithm of IUP-UB. For the IUP-UB algorithm the location (i.e., latitude and longitude) of a given target region is input, whereas the SRON and Kayrros algorithm determine the location of the potential emission source by analysing the entire global satellite data set. Therefore, a spatial collocation criterium has been defined and for the comparison results shown here a maximum distance of 20 km has been selected (corresponding to approximately 3-4 ground pixels of TROPOMI). This comparison is performed on one year of data (year 2021).

5.1.2 Results

The comparison shows that, as expected, the IUP-UB Cross-Sectional-Flux (CSF) algorithm provides typically the largest number of methane emission estimates for a given target location (see **Tab. 5.1.2-1**). This is because this algorithm (in contrast to the other 2 algorithms) is not based on emission plume identification. Instead, only a small set of selection criteria is used in order to decide if the satellite data can be used for emission estimation or not: Data coverage must be dense enough (e.g., > 90% of the area used for the CSF analysis needs to be covered by “good” (e.g., cloud free) data independent of the spatial pattern of the methane concentration field) and wind speed should not be too low (> 1 m/s). If these criteria are met than the quality flag (QF) attached to the corresponding emission estimate is set to “good” (QF=0) (otherwise it is set to 1). Consequently, estimated emissions using IUP-UB’s CSF

algorithm may scatter around zero in case no methane plume is present (because of the noisy XCH_4 retrievals). The smallest number of emission estimates are provided by the Kayrros algorithm, because of the combined effect of the computer-vision approach used for plume detection and of the inversion algorithm based on detailed modelling, which is challenging due to the required good knowledge of meteorological information (e.g., wind field), issues related to turbulence, required match between modelled and observed plume, etc. The number of estimates provided by SRON is in between the number of estimates of Kayrros and IUP-UB. For a detailed comparison of the global SRON and Kayrros emission estimates see **Sect. 6**, where it is also shown that the SRON algorithm produces a significantly larger number of emission estimates compared to the Kayrros algorithm and additional explanations are given why this is the case.

Primarily due to the low number of Kayrros estimates and the fact that often only one algorithm provided an estimate for a given target location and satellite overpass, the number of collocations is small. At 3 of the 10 target locations no collocations have been found (namely the targets in Upper Silesia, Karaturun East and Hambach, see **Tab. 5.1.2-1**).

For the 7 other locations where collocations exist, the individual emission estimates often differ significantly, e.g., between IUP-UB and SRON, which are the 2 algorithms with by far the largest number of collocations (see **Fig. 5.1.2-1**). The agreement between Kayrros and SRON is often much better (differences are typically some 10%) but this finding is based on only 10 collocations (in contrast, 65 collocations have been identified for IUP-UB and SRON). For 3 of the 7 target locations where collocations exist, the median of the ratio of the IUP-UB and the SRON estimates is close to 1 indicating (on average) good agreement but for the other 4 of the 7 cases the median of the ratio is larger than 2.

Target ID Short name	Emissions (2021) All in 2021; Mean ± Stddev [t/h] (Ndata)			Differences (#): Absolute [t/h] (Ncoloc) Relative Mean ± Stddev [%] (R)			Comments
	SRON	Kayrros	IUP-UB	SRO - Kay	IUP - Kay	IUP - SRO	
FM-01 Hassi_Messaoud	27 ± 14 (18)	29 ± 6 (4)	17 ± 54 (147)	- -	- -	8 ± 26 (12) 4 ± 73 (0.64)	
FM-02 Korpezhe	50 ± 26 (33)	48 ± 23 (3)	56 ± 48 (29)	- -	- -	50 ± 38 (8) 74 ± 48 (-0.25)	
FM-03 Valdemingomez	13 ± 11 (18)	- (0)	23 ± 30 (49)	- -	- -	8 ± 24 (11) 75 ± 194 (0.46)	
FM-04 Norte_III	64 ± 30 (28)	63 ± 21 (9)	78 ± 39 (17)	31 ± 36 (5) 36 ± 44 (0.32)	- -	13 ± 37 (10) 15 ± 42 (0.39)	
FM-05 Upper_Silesia_Zory	25 ± 16 (4)	- (0)	82 ± 30 (3)	- -	- -	- -	
FM-06 Hail_Creek	24 ± 7 (8)	18 (1)	28 ± 32 (24)	- -	- -	21 ± 21 (5) 26 ± 92 (0.57)	
FM-07 Permian_Pecos	27 ± 12 (8)	- (0)	56 ± 61 (124)	- -	- -	66 ± 33 (5) 105 ± 13 (0.94)	
FM-08 Shanxi_Qinshui	91 ± 44 (25)	69 (1)	233 ± 139 (28)	- -	- -	145 ± 125 (14) 57 ± 89 (0.76)	
FM-09 Karaturun_East	- (0)	- (0)	-17 ± 28 (7)	- -	- -	- -	No plumes in 2021
FM-10 Hambach	- (0)	- (0)	33 ± 30 (12)	- -	- -	- -	No plumes in 2021

Ndata: SRON & Kayrros: Number of plumes; IUP (WFMDv1.8, CSFv1.6): Number of overpasses with dense data coverage

(#) Only reported if more than 2 colocations (Ncoloc > 2)

Relative difference [%] := (a-b)/mean (example: a=2, b=1 -> 67%)

Tab. 5.1.2-1: *Summary of the target region comparison. Columns in "Emissions (2021)" list the mean +/- standard deviation of estimated emissions computed independently for the three algorithms including number of data (Ndata). Emission differences are listed in the following 3 columns but only if the number of pairs of collocated data (Ncoloc) is larger than 2.*

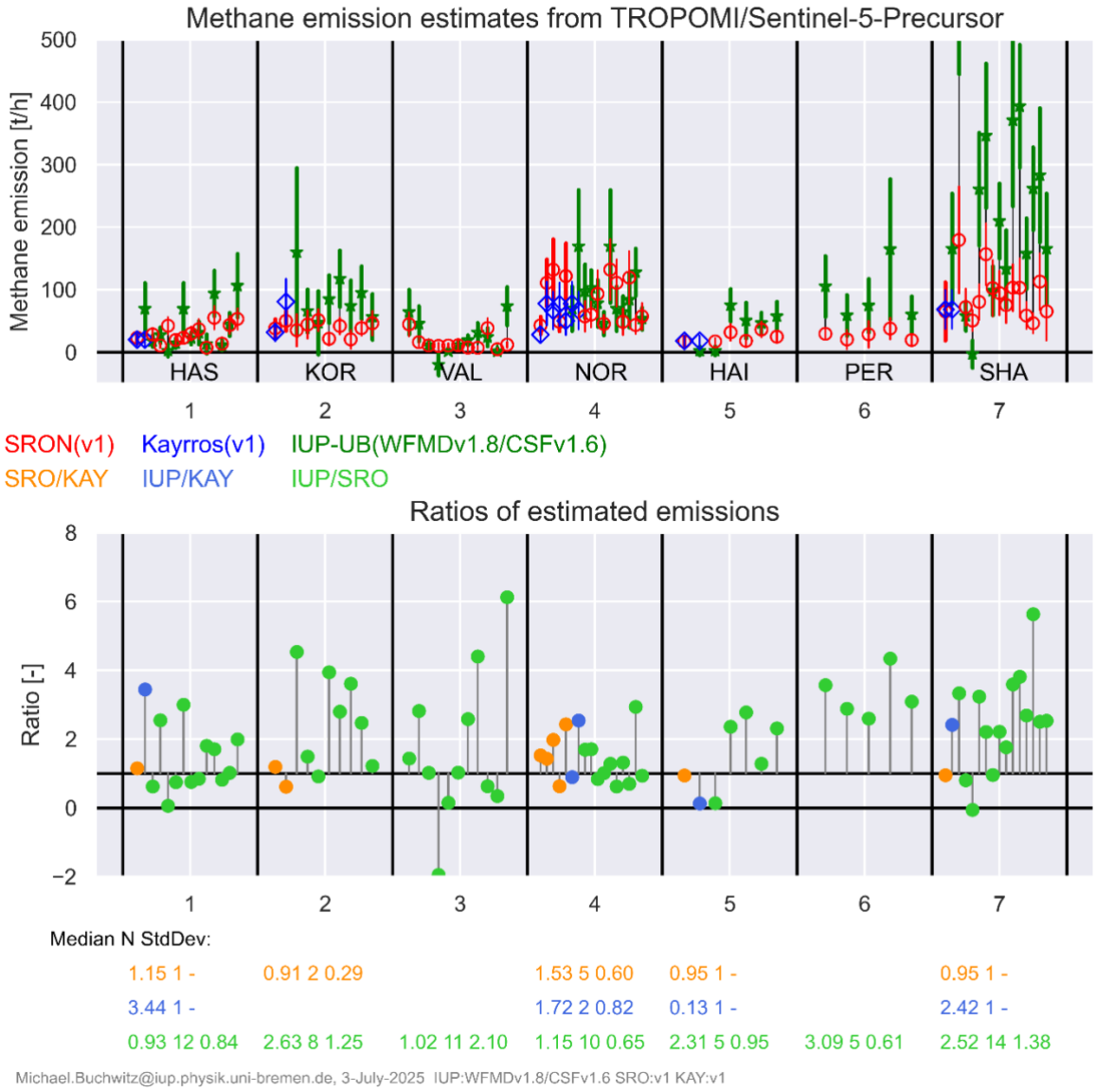


Fig. 5.1.2-1: *Comparison of estimated emissions for target locations with collocated emissions (7 from 10). Top: Estimated emission. Bottom: Ratios of estimated emissions. For each of the 7 target regions the following numerical quantities are listed: the median of the ratios, the number of collocations N and (if N > 1) then standard deviation of the ratios.*

Combining the IUP-UB CSF v1.6 and SRON collocations for all 7 target locations shows that the IUP-UB estimates are on average about a factor of 2 larger compared to SRON (see **Fig. 5.1.2-2**).

Several assessments have been carried out to explain the relative overestimation of the IUP-UB CSF v1.6 emission estimates compared to SRON at some of the target locations. It has been identified that the different Level 2 input data are not the driver for these differences, and it is argued that also the wind speed used by IUP-UB is unlikely the reason for a factor of 2 or larger differences. IUP-UB uses for TROPOMI the Planetary Boundary Layer (PBL) average of the wind speed, which is in line with Varon et al., 2019, who found that the PBL average wind speed is appropriate for TROPOMI CSF-based emission estimates. Also, initial assessments based on simulated plumes have not shown that there is an obvious problem with the IUP-UB CSF algorithm that leads, for example, to a significant overestimation of the estimated emissions.

These activities also resulted in a new version 1.7 of the IUP-UB CSF algorithm. The main differences of v1.7 compared to v1.6 are a modified background correction and a waiver of the 3-hour plume length limit used for v1.6. The waiver of the 3-hour limit typically results in a reduction of the emission estimates as typically the computed fluxes through the cross-sections are lower for cross-sections further away. In this context it is also important to know how the reported emissions need to be interpreted. For v1.6, with the 3-hour limit implemented, the reported emission referred to the average emission in the period 3 hours before the satellite overpass to the time of the satellite overpass. For v1.7 this averaging time is not constant anymore but depends on wind speed. For an emission source emitting at a constant rate both approaches lead (ideally) to identical results but this is not true for sources with time dependent emission.

As shown in **Fig. 5.1.2-3**, v1.7 results in much better agreement with the SRON estimates indicating that the implemented changes to the IUP-UB CSF algorithm are related to aspects responsible for at least large parts of the observed differences.

Due to the limited number of collocations with the Kayrros estimates at the selected 10 target regions, the target region comparison focused on comparisons between the IUP-UB and SRON algorithms. For a more meaningful comparison of the SRON and the Kayrros algorithms, please see the following global comparison results.

5.1.3 Potential future research

So far only a limited amount of data has been processed and analyzed with CSF v1.7 and further studies are needed to find out if v1.7 is better than v1.6 also for cases not yet investigated.

Other aspects include further assessments related to simulated plumes and validation related aspects which are addressed in WP 200.

All these activities will result in additional comparisons including assessments / sensitivity studies using modified algorithms (e.g., using perturbations of potentially critical parameters), where appropriate.

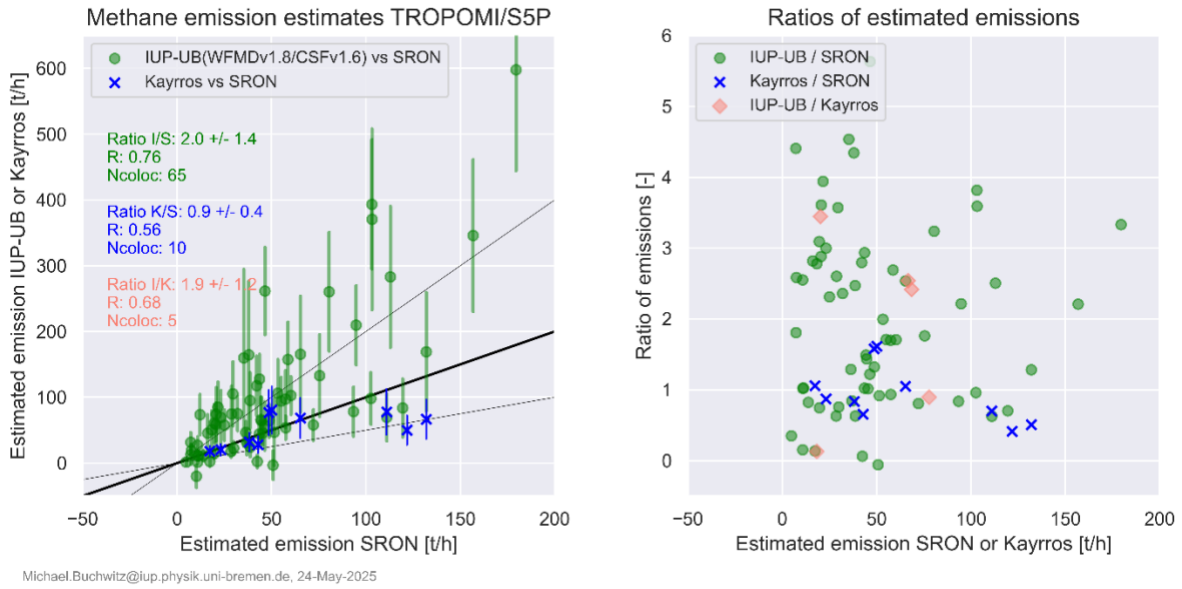


Fig. 5.1.2-2: *Left: Comparison of collocated emission estimates (green: IUP-UB (CSF v1.6) & SRON; blue: Kayrros & SRON; light red: IUP-UB and Kayrros (only the numerical results are shown for this pair of algorithms)). Listed are the numerical values of the following quantities: ratio of emissions in terms of mean value and standard deviation, linear correlation coefficient R, number of collocations Ncoloc. Right: corresponding ratios of emissions.*

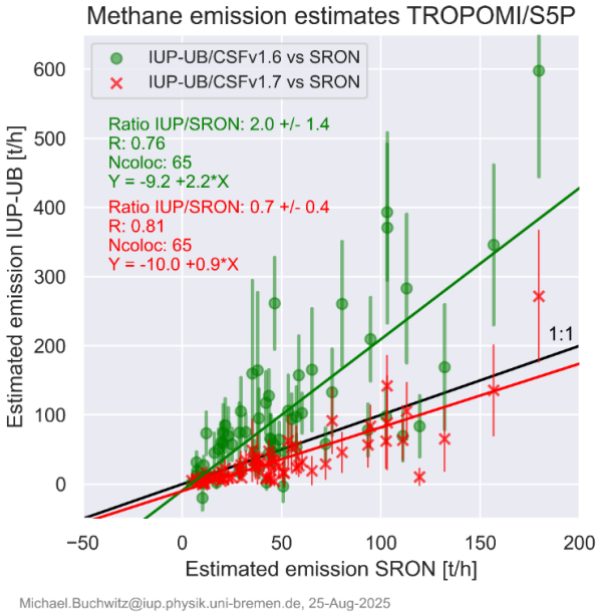


Fig. 5.1.2-3: *Similar as* ***Figure 5.1.2-2 (left)*** *but limited to the IUP-UB and SRON comparison and with IUP-UB CSF v1.7 results added (in red) in addition to the CSF v1.6 results (in green).*

5.2 Global intercomparison results

5.2.1 Purpose and approach

The purpose of this intercomparison is to extend the comparisons presented in the previous section using all globally obtained emission estimates based on TROPOMI/S5P data during the year 2021. Both the algorithm from SRON and the Kayrros algorithm can analyze all TROPOMI data with respect to detection of localized emission sources including estimation of the location of the plume origin, i.e., the approximate location of the emission source. This contrasts with the current version of the IUP-UB algorithm, where the emission source location is a required input for emission quantification. The results in this section are therefore limited to comparisons between the SRON and Kayrros algorithms with respect to detection and quantification.

5.2.2 Results

Global datasets of methane emission plumes detected in TROPOMI data have been reported by Kayrros and SRON for the year 2021. Both datasets have detections covering the main methane emitting regions (e.g., Permian basin, West coast of Turkmenistan, Shanxi coal mines, etc.), with SRON also covering additional locations (e.g. Casablanca, Istanbul, and coal mines in Poland) consistent with a three-times higher number of detections (3233 SRON versus 1060 Kayrros, see **Fig. 5.2.2-1**).

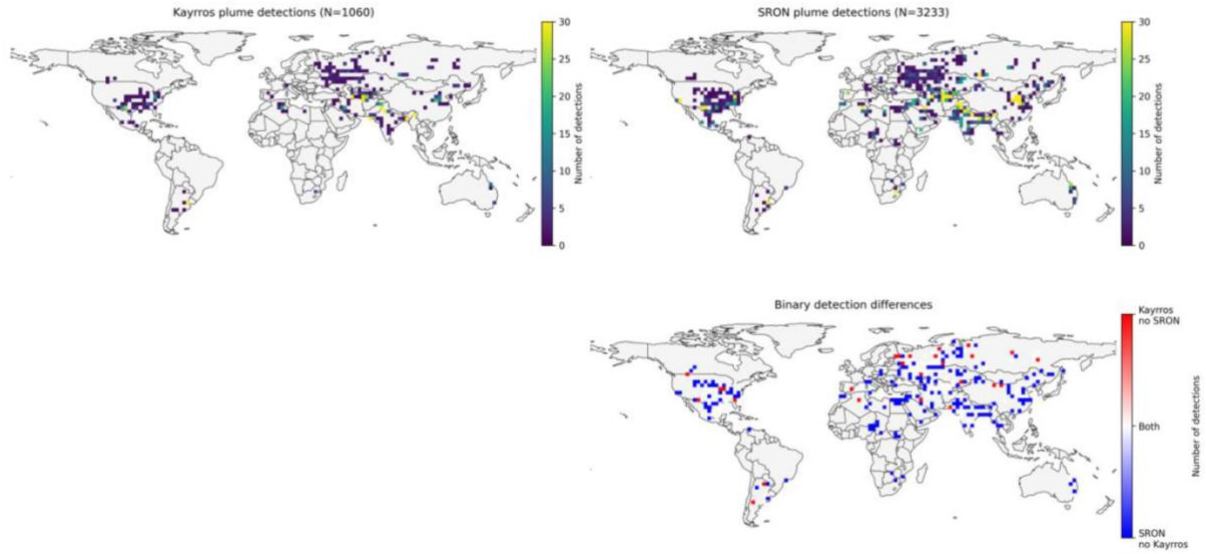


Figure 5.2.2-1: *Spatial gridded (2.5° x 2.5°) distribution maps of methane plumes detected in TROPOMI data for the year 2021 by Kayrros (top left), SRON (top right) and their binary difference (bottom right), highlighting where at least one plume has been detected by one group only.*

Plumes from both datasets have been matched based on the comparison of their reported plume masks. The Kayrros – SRON matching plumes (N=643, with 306 that include emission quantifications in both as not all detected plumes were quantified by Kayrros) show excellent agreement (R = 0.62, low mean bias) in estimated methane emission rates, considering that Kayrros and SRON quantification methods are strictly different (except for the input TROPOMI data). However, the scatter of Kayrros – SRON differences is quite significant, with a 53% relative difference standard deviation (see **Fig. 5.2.2-2 (top left)**), with 77% agreeing within uncertainties.

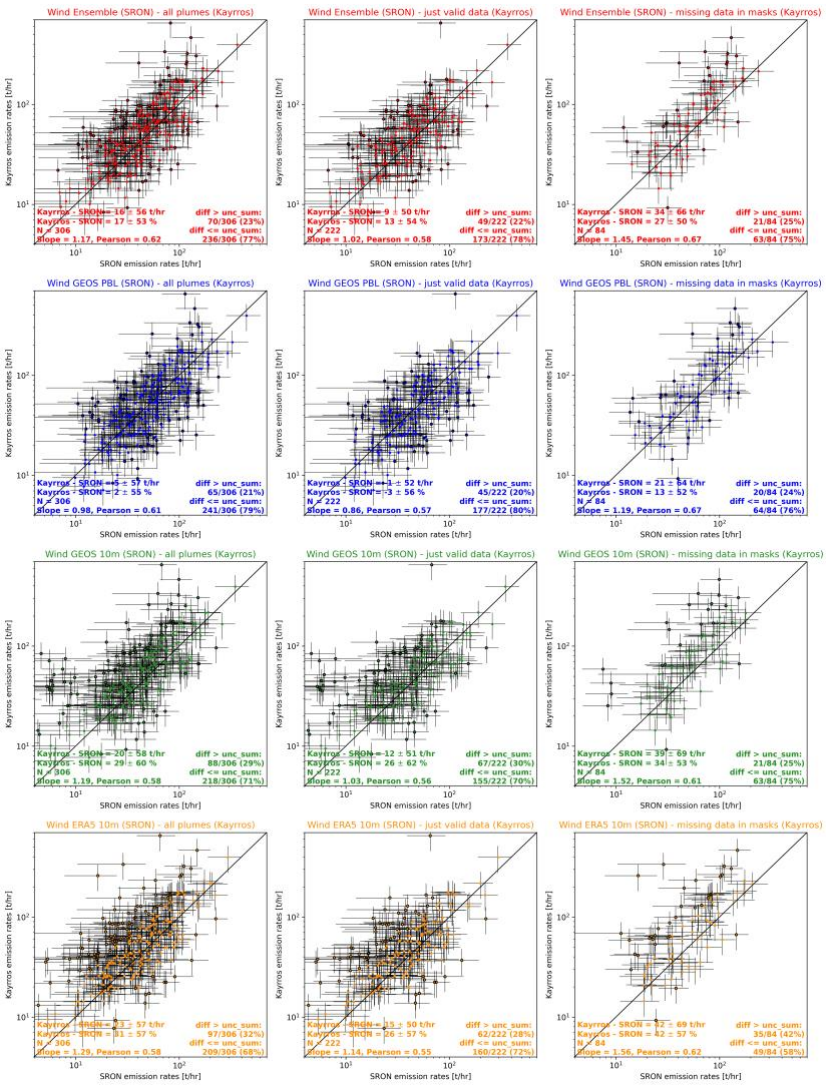


Figure 5.2.2-2: *Comparison of Kayrros and SRON emission rates for all one-to-one matches and quantified plumes (left, N=306), for all plumes with only valid data in their masks (center, N=222) and for plumes with missing pixels in Kayrros plume masks (right, N=84), using the full SRON wind ensemble result (top line, red) and using just GEOS PBL averaged (2nd line, blue), GEOS-FP 10 m (3rd line, green) and ERA5 10 m (bottom, orange) wind speeds. One-to-one matching plumes highlighted with black outlines do not have overlapping uncertainty ranges.*

Most of the plume-to-plume emission difference scatter is shown to be explained by the difference between Kayrros and SRON emission quantification methods as shown in **Fig. 5.2.2-3** where in the right figure the SRON emission quantification method was used on the Kayrros plumes

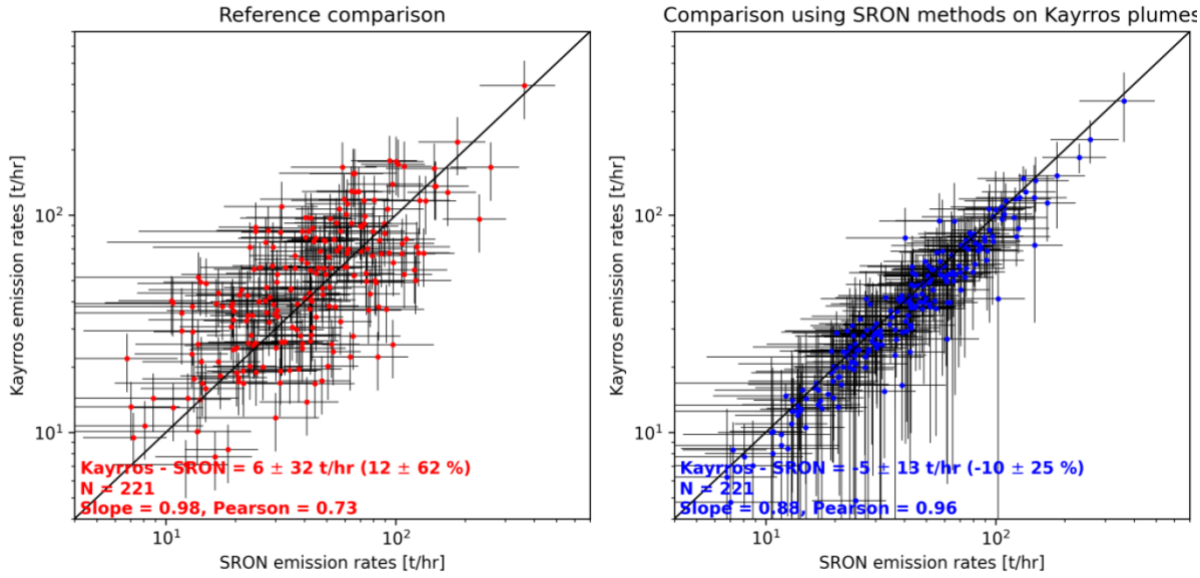


Figure 5.2.2-3: *Kayrros – SRON emission rate comparison for Kayrros emission rates quantified using the Kayrros quantification method (left) and for Kayrros emission rates quantified by using the SRON quantification method on the Kayrros-detected plumes (plume mask + estimated background XCH4 values). Inset percentages are provided relative to the averaged SRON plume emission rate.*

Examining the 386 plumes detected by Kayrros but not reported by SRON, it has been identified that filtering (TROPOMI pixel filters and human labelling of detections) employed by the SRON TROPOMI plume detection framework can explain the large majority of these non-detected or non-reported plumes. Based on the features that can be computed from the plumes detected by SRON, the 2568 plumes not reported by Kayrros in 2021 appear to mainly be the shorter and less intense plumes, compared to the Kayrros – SRON matching plumes.

5.2.3 Potential future research

In order to further interpret the intercomparisons it would be good to use existing model simulations and/or to produce additional ones generating XCH_4 maps based on known emissions and to apply the emission estimation algorithms to these data to find out how well the estimated emissions agree with the known emissions as used as input for the model simulations. This may help to further improve the algorithms.

6 Intercomparison results for Hyperspectral Imager (HI)

This section summarizes the intercomparison results for the HI instruments. All details are reported in ANNEX B (see **Sect. 9**).

The intercomparison has been split according to four different parts, each addressing a specific aspect:

- Part 1: Plume identification and attribution to source
- Part 2: Assessment of methane concentration retrievals used for detection
- Part 3: General methane emission quantification assessment
- Part 4: Detailed methane emission quantification assessment

6.1 Part 1: Plume identification and attribution to source

6.1.1 Purpose and approach

An initial step of several of the used algorithms is methane plume identification and attribution to the corresponding methane emission source. Plume detection, i.e., the identification of locally elevated atmospheric methane concentrations due to an emission source, is used by the algorithms of SRON, UPV, and Kayrros but not the CSF algorithm of IUP-UB. Therefore, SRON, UPV, and Kayrros contributed to these intercomparisons.

The comparison focussed on 2 scenes and 2 sensors: An EnMAP image capturing an area in the Shanxi coal mines in China and an EMIT acquisition of an oil and gas (O&G) region in Turkmenistan. The EnMAP scene exhibits a relatively dark albedo and a heterogeneous surface, while the EMIT one is bright and homogeneous. Both acquisitions contain numerous plumes originating from a wide range of emission rate values.

6.1.2 Results

Regarding the emission detection performance, we found agreement among groups in ideal scenarios (bright and homogeneous areas), where differences were only found in blob-like emissions at background level, i.e. emissions that do not follow a typical plume shape and are at retrieval background levels. The advantages of machine learning-based models in detection could not be correctly assessed due to the low number of samples in our tests. Under less ideal scenarios, emission detection seems to be highly dependent on the criteria individual groups use. A more conservative criteria generally leads to a lower number of detections, while a more

relaxed criteria increase the number of detections while assuming a higher risk of false positives (though very few false positives were found).

Our analysis shows that groups specific criteria for plume delineation can lead to important differences, which in turn could have a significant impact in quantification but can also be compensated for by different calibrations. Some of the different criteria considerations are whether to allow not connected clusters to be part of the same plume, the consistency of the plume masking with the quantification calibration, and human-derived errors linked to distinguish background from plume pixels.

Regarding potential emitter locations and plume delineations, **Fig. 6.1.2-1** shows results for the Turkmenistan and China scenes from SRON, Kayrros and UPV. As can be seen, the extends of the identified plume areas can be quite different.

As shown in **Fig. 6.1.2-2**, there is good to reasonable agreement in the number of identified plumes.

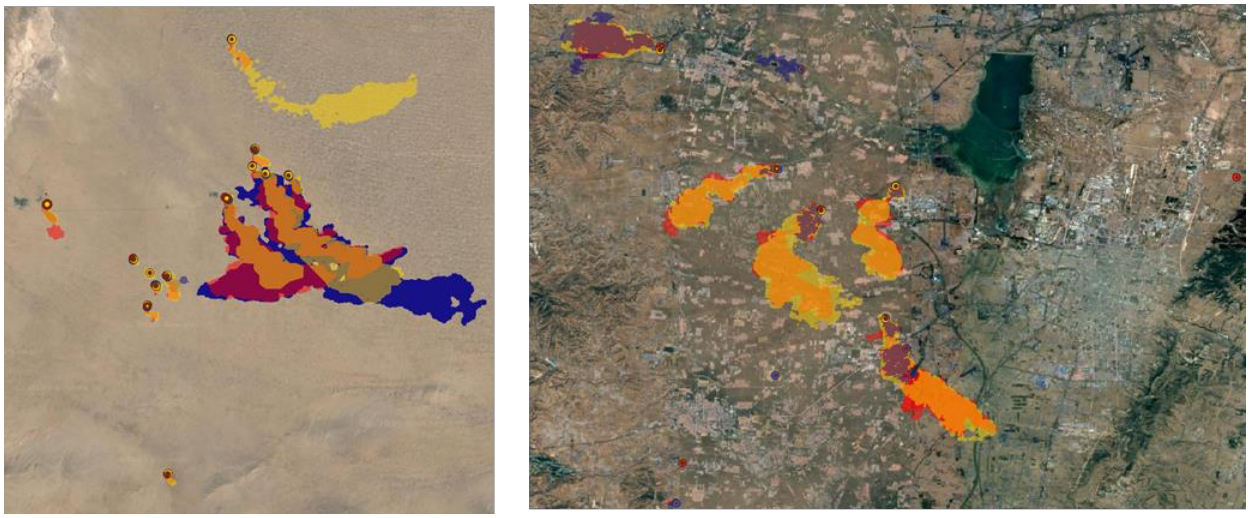


Figure 6.1.2-1: *Potential emitter locations (points) and plume delineations of the SRON (blue), Kayrros (yellow), and UPV (red) groups for the EMIT scene in Turkmenistan (left) and the EnMAP scene in China (right).*

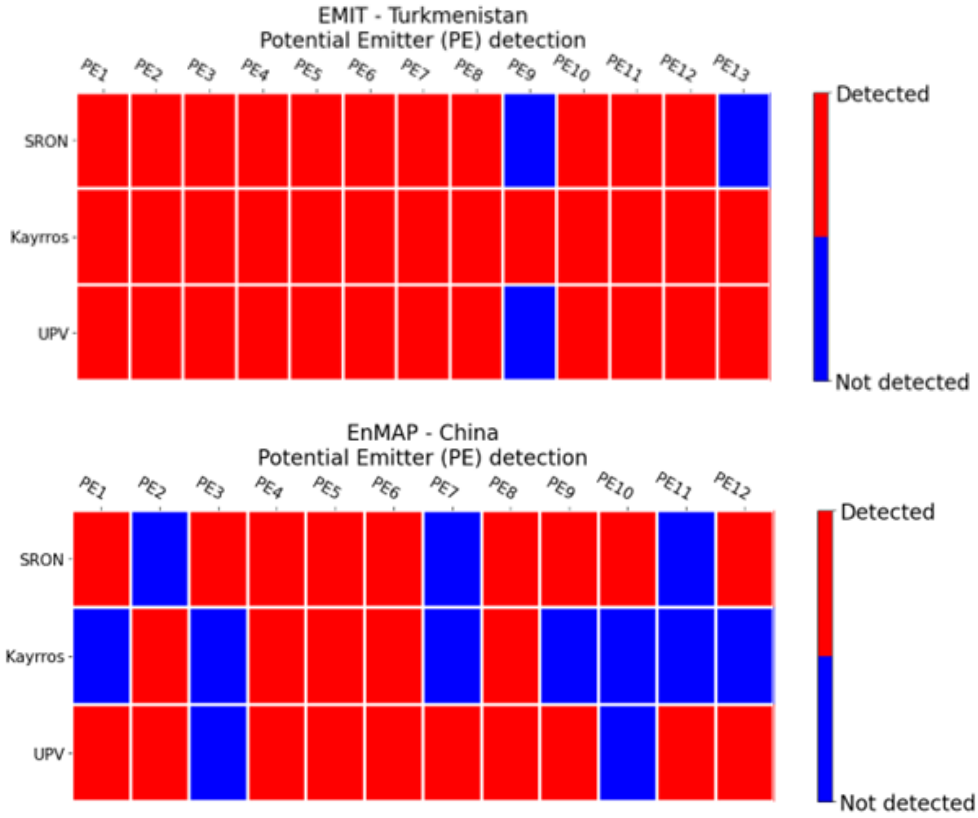


Figure 6.1.2-2: *Potential emitter detection matrix for the Turkmenistan / EMIT scene (top) and the China / EnMAP scene (bottom).*

6.1.3 Potential future research

No specific future research topics have been identified but an extension to more target regions would very likely be beneficial for further improvements of the used algorithms.

6.2 Part 2: Assessment of methane concentration retrievals used for detection

6.2.1 Purpose and approach

Purpose of this part of the intercomparison was to investigate to what extent the retrieved atmospheric methane maps agree focussing also on regions assumed not to have any major methane sources. This is relevant to see to what extent the methane images are affected by systematic artifact and by noise, which have an impact in the detection and quantification of real methane emissions. Nine scenes covering different conditions in terms of surface brightness or heterogeneity have been identified for this comparison for EnMAP, EMIT and PRISMA. All four HI groups, i.e., SRON, UPV, Kayrros and IUP-UB, participated in this exercise.

6.2.2 Results

All the groups detection retrievals showed similar scene and instrument features, showing consistency. Moreover, we found that the selection of the retrieval spectral window has an important impact in detectability when applying matched filter methods. Combo-MF, a matched filter flavour, exploits this point reducing the occurrence of retrieval artifacts, which is helpful to filter them out of the plume mask. This is especially important for the IME quantification, since plume delineation is needed. We also found that other factors such as a proper radiance quality filtering can lead to significant changes, since radiance 'bad pixels' interfere with scene-derived parameters used by the matched filter methods.

Figure 6.2.2-1 shows methane retrieval from the 4 algorithms/groups as obtained from EnMAP radiances over the Anna Creek area in Australia. See Annex B for a detailed discussion, where it is also highlighted that some of the differences are due to slight offsets (see **Figure 6.2.2-2**) between the 4 data sets (which are only of very minor importance for emission quantification).

Overall, the agreement of the methane enhancements in terms of mean values and their scatter is very good, as shown in **Figure 6.2.2-3**.

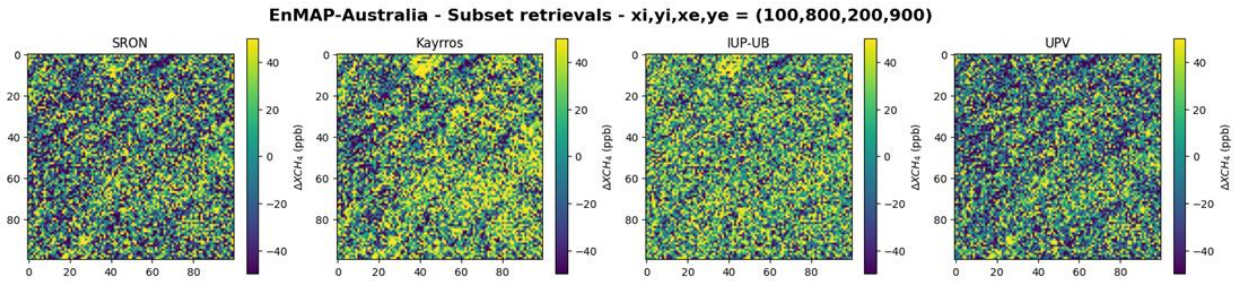


Figure 6.2.2-1: *Methane retrievals from the selected subset of the EnMAP acquisition capturing an Anna Creek (Australia) area of the (starting from the left) SRON, Kayrros, IUP-UB, and UPV groups.*

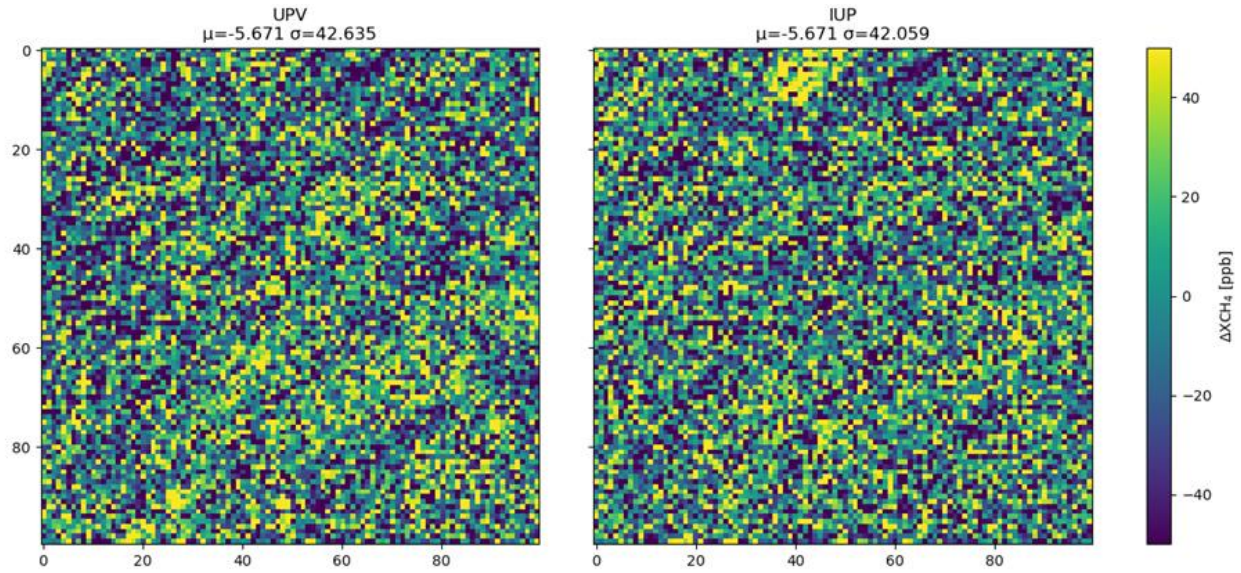


Figure 6.2.2-2: *Similar as* ***Figure 6.2.2-1*** *but limited to IUP-UB (right) and UPV (left) but with a small offset added to the IUP-UB methane enhancement image.*

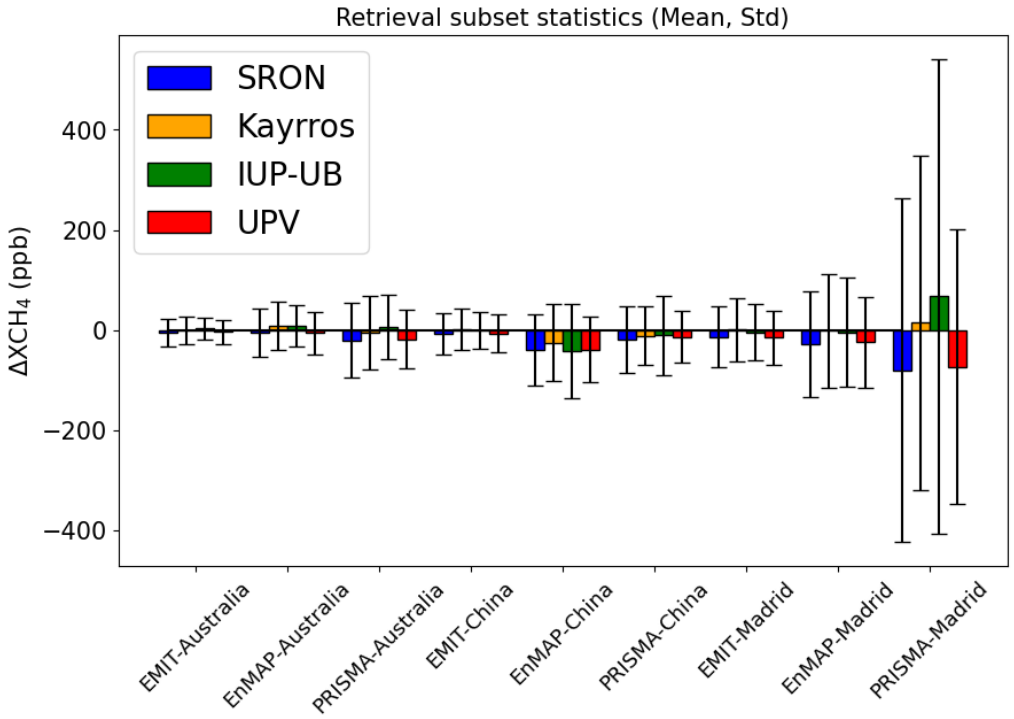


Figure 6.2.2-3: *Mean (bars) and standard deviation (error bars) values related to the SRON (blue), Kayrros (yellow), IUP-UB (green), and UPV (red) groups of all the retrieval subsets used for the assessment of methane concentration retrievals used for detection activity.*

6.2.3 Potential future research

No specific future research topics have been identified for this part of the comparison.

6.3 Part 3: General methane emission quantification assessment

6.3.1 Purpose and approach

The main goal of this part of the intercomparison was to obtain a first overview of how well the estimated emissions agree or disagree. For this purpose, 12 satellite acquisitions have been carefully selected containing methane emission plumes in images obtained from EnMAP, PRISMA and EMIT covering a wide range of scenarios. All four HI groups, i.e., SRON, UPV, Kayrros and IUP-UB, participated in this exercise.

6.3.2 Results

The selection of the wind speed value has a great impact on the quantification results. In some cases, we found that groups selected wind speed values that differ up to 2 m/s for the same plume case, which shows potential high disagreement.

These differences arise from the specific selection of values from GEOS-FP or ERA5 as data source, and the corresponding postprocessing.

Regarding the Q values reported by the teams, we also observe a certain degree of variability, which cannot only be explained by wind speed variability. The use of different methane retrievals and quantification also play an important role. In some cases, we find remarkable differences, such as groups having a Q 8 times higher than another (see **Fig. 6.3.2-1**). In addition, there is a trend of IUP-UB showing systematically lower Q values than the rest of the groups for most cases. This point needs further research. Nevertheless, except for a few cases, we generally find that group estimates agree within the uncertainty bars.

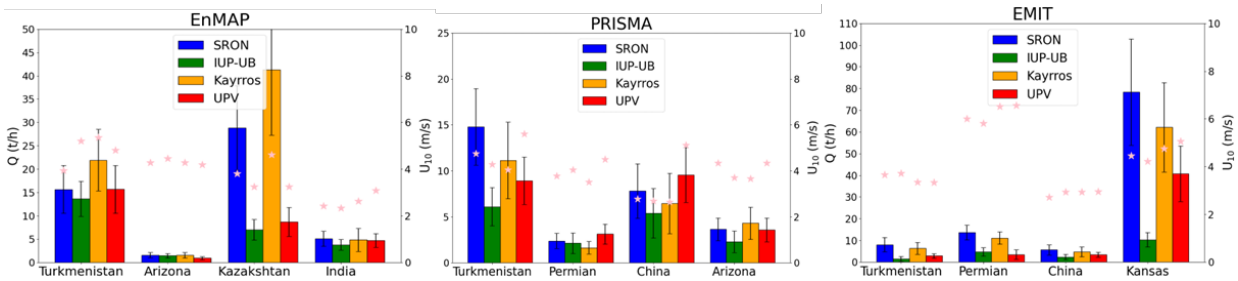


Figure 6.3.2-1: *Flux rate estimations (left y-axis, bars) and U10 values (pink stars, right y-axis) reported by SRON (blue), Kayrros (yellow), IUP-UB (green), and UPV (red) teams for several methane emission cases obtained using EnMAP (left), PRISMA (center), and EMIT (right) data.*

6.3.3 Potential future research

We still cannot explain why most IUP-UB quantification results were systematically lower than those from the rest of the groups. IUP-UB is the only group using the CSF method, while the other groups used the IME method. Therefore, these systematic differences are likely to be found in that direction. However, further research is needed to understand this point.

6.4 Part 4: Detailed methane emission quantification assessment

6.4.1 Purpose and approach

This Part 4 is an extension of Part 3 focussing on further detailed investigations for selected scenarios, i.e., on a sub-set of the scenarios studied in Part 3. This includes also the production of new results obtained by varying parts of the algorithms in terms of the used wind speed and the underlying input atmospheric methane image. 3 scenes have been investigated in detail: An EnMAP and a PRISMA scene in Arizona (where also comparisons from past controlled release experiments were available) and an EMIT scene in Kansas. All four HI groups, i.e., SRON, UPV, Kayrros and IUP-UB, participated in this exercise.

6.4.2 Results

As shown in Part 3 (**Sect. 6.3.2**) especially the EMIT / Kansas scene showed large differences of the estimated methane emissions, with IUP-UB reporting significantly lower emissions compared to the other algorithms. **Figure 6.4.2-1** shows the retrieved methane enhancements for the EMIT / Kansas scene as obtained from the 4 algorithms and **Figure 6.4.2-2** shows corresponding scatter plots of the enhancements as retrieved from Kayrros (left), IUP-UB (middle) and UPV (right) compared to SRON (x-axis). The largest values and also the linear fit (red line) suggest that the agreement is best between UPV and SRON. Based on that one would expect that these two algorithms also agree best in terms of the methane emission estimates (and as shown in **Fig. 6.4.2-5** top left, this is in fact the case).

After a comprehensive analysis of the emission estimates of 3 plume cases we found that there can be large differences even when using the same quantification method due to the decisions taken within the method itself. For example, **Fig. 6.4.2-3** bottom right panel shows that the estimated emissions agree very well, when the same methane enhancement images ("common retrieval") and same wind speeds are used but **Figs. 6.4.2-4** and **6.4.2-5** show that this is not always the case.

As for the methane retrievals used for quantification, we observed that some groups exhibit different levels of retrieval plume values and we found that most discrepancies were caused by the selection of the unit methane absorption spectrum (i.e., K spectrum, see **Fig. 6.4.2-6**), which plays an important role in the matched filter performance. The methane concentration enhancement range used to fit this spectrum and the use or not use of radiometric spectra corrections (Gorroño et al., 2023) explain a large fraction of these differences.

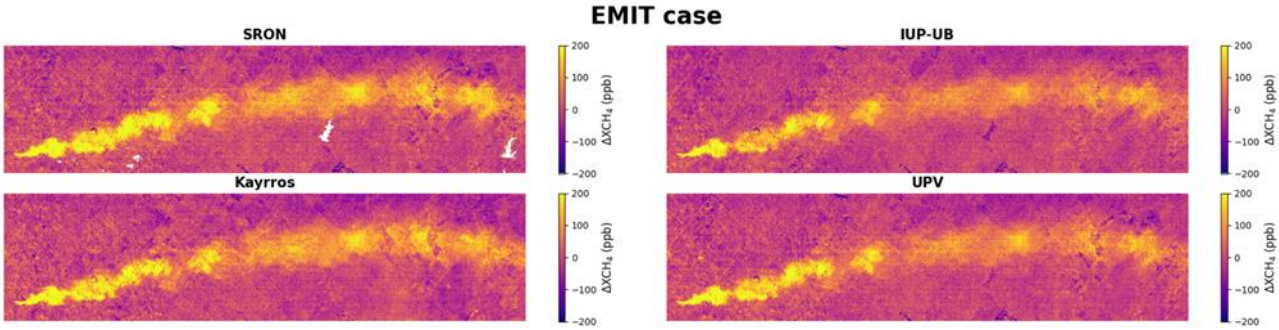


Figure 6.4.2-1: *Subsets of methane concentration enhancement maps from each group capturing the plume characterized with EMIT data.*

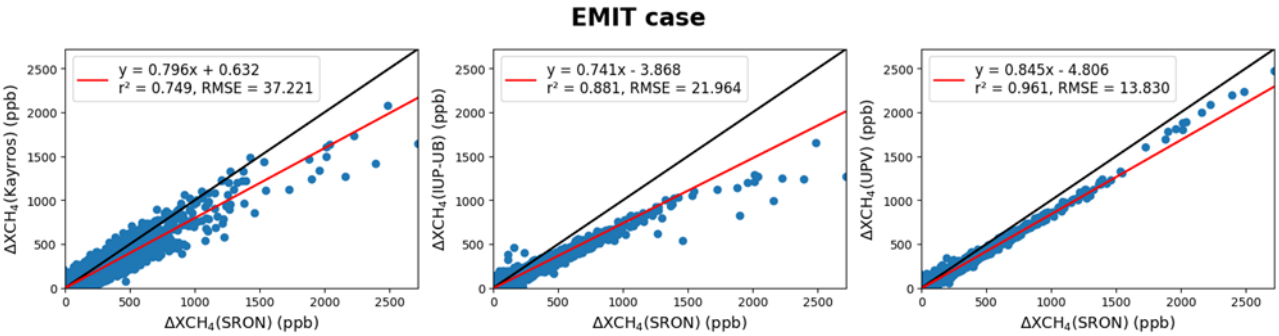


Figure 6.4.2-2: *Scatter plots comparing the retrieval values from SRON to those from Kayrros (left), IUP-UB (center), and UPV (right) for the plume captured with EMIT data.*

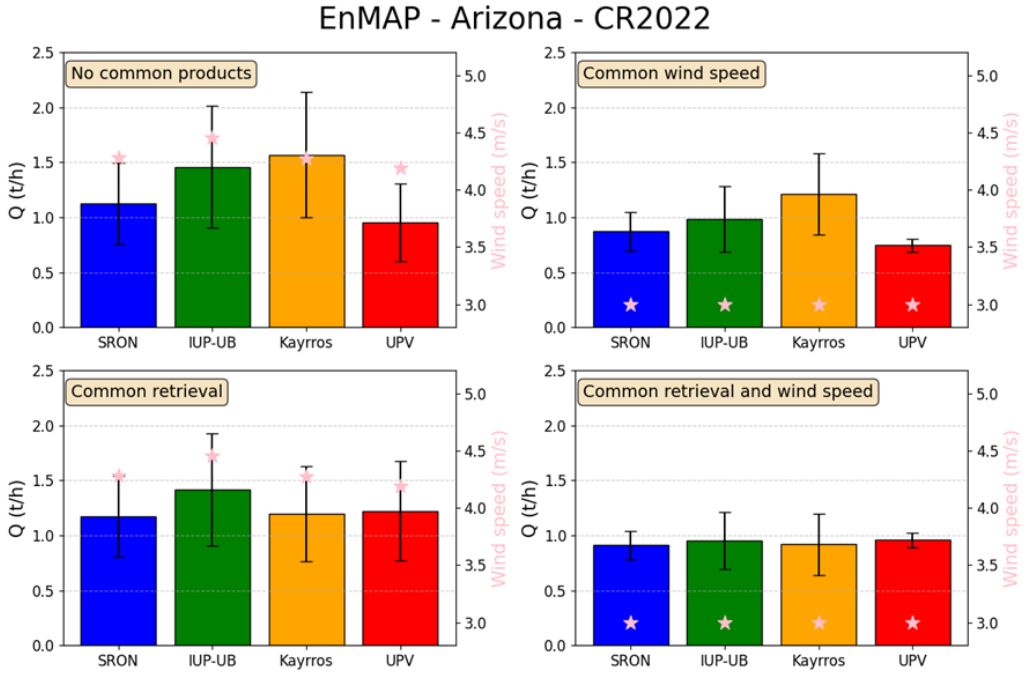


Figure 6.4.2-3: *Emission rate estimations (bar charts) with their related errors (error bars) for the 4 different quantification cases along the used wind speed values (pink stars) for a plume located in Arizona and characterized using EnMAP data.*

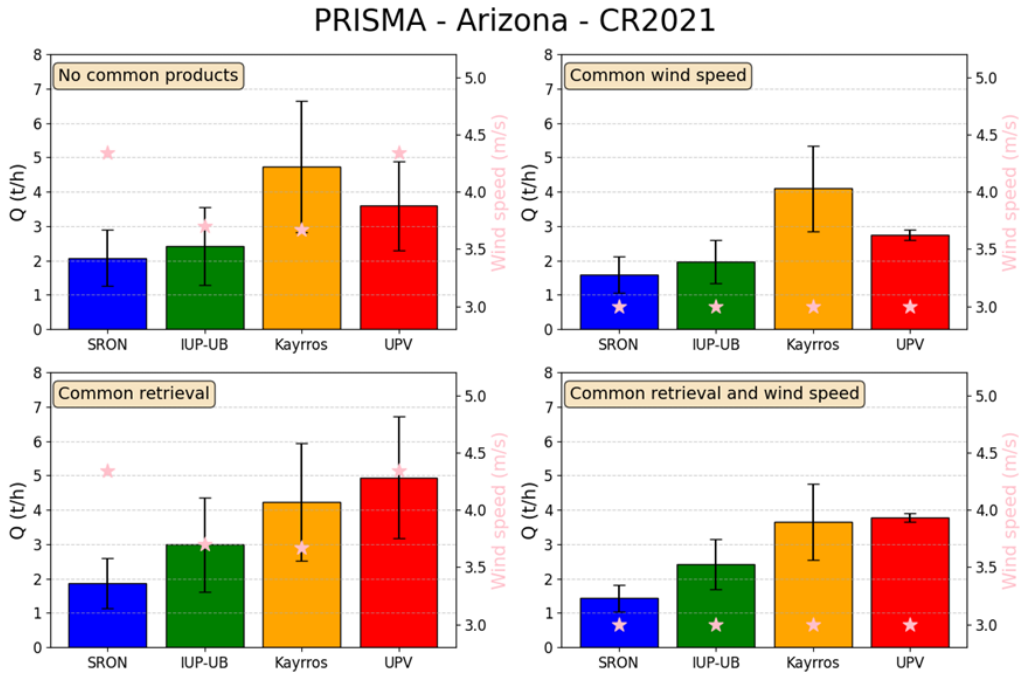


Figure 6.4.2-4: *Similar as* ***Fig. 6.4.2-3*** *but for the PRISMA Arizona scene.*

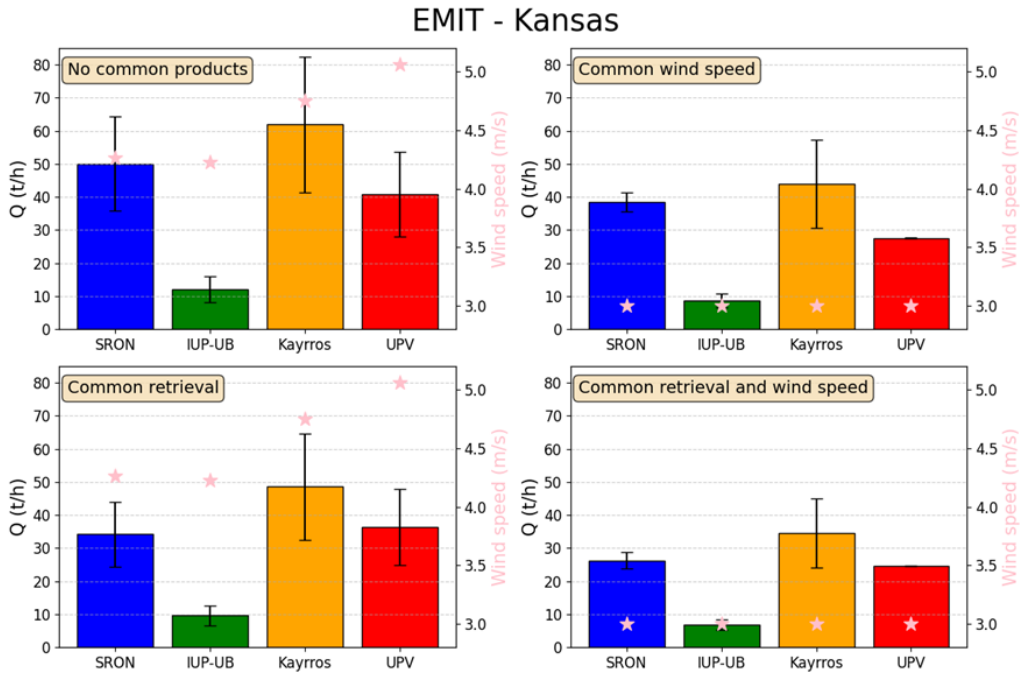


Figure 6.4.2-5: *Similar as* ***Fig. 6.4.2-3*** *but for the EMIT Kansas scene.*

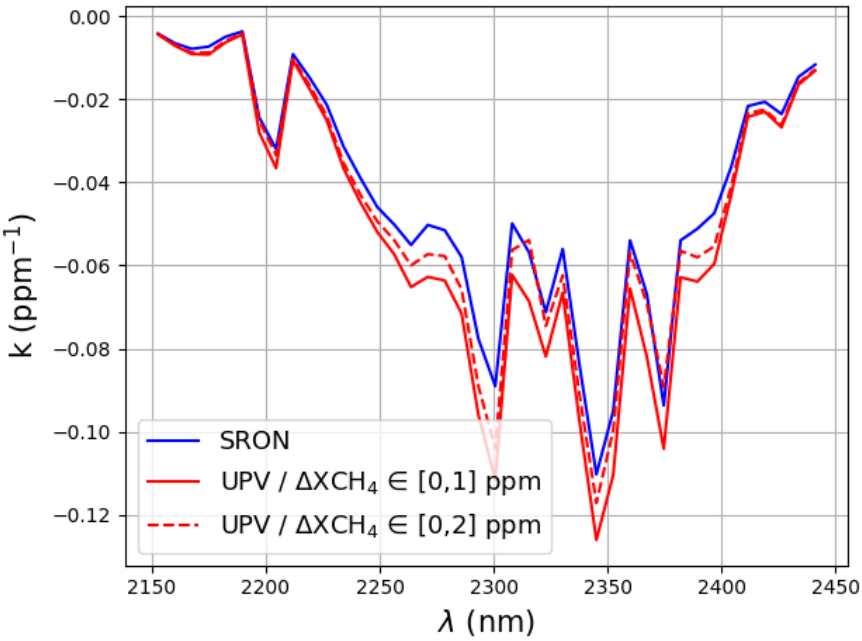


Figure 6.4.2-6: *K spectra used by SRON (blue) and UPV (red) for the EMIT case. UPV considers a K derived from ΔXCH_4 values between 0 and 1 ppm (solid line) and between 0 and 2 ppm (dashed line).*

Regarding the IME method, the selection of the plume mask plays an important role in the flux estimate. Similarly, the plume coverage extension is an important factor in the CSF method.

We also found that there can be consistency between estimates coming from different quantification methods, such as in the PRISMA and EnMAP cases after setting the same retrievals and wind speed values. Moreover, there are generally non-negligible changes after fixing the retrieval, which matches with the different selection of the unit methane absorption spectra from groups. In addition, we did not see wind speed as a major contributor to discrepancy in these 3 cases explored for the detailed methane emission quantification assessment (differences only up to ~ 1 m/s), unlike differences up to ~ 2 m/s that were found in other cases from the general methane emission quantification assessment (**Sect. 6.3**). Overall, most major discrepancy factors have been detected, which reveals the success of the analysis.

6.4.3 Potential future research

The following topics have been identified:

- To carry out a more in-depth analysis on how Jacobian K (i.e, the change of the radiance (or its logarithm) for a +1 ppm methane concentration increase) is

obtained. Consideration about Radiative Transfer Models (RTMs) and how the Look-Up-Tables (LUTs) are built might explain remaining differences.

- To carry out a comprehensive analysis on cases where there is a higher difference in used wind speed values (e.g., PRISMA China case. In this manner, we would be able to better assess to what extent the quantification results are impacted by wind speed differences).
- To carry out a comprehensive analysis on a significantly larger number of cases to better assess the main sources of discrepancies. The 3 cases investigated are not enough to draw robust conclusions.
- Despite additional investigations, we still cannot explain why the IUP-UB quantification results for the EMIT Kansas scene were systematically lower than the results of the other groups.
- To link the conclusions from what was obtained in this work package to others, such as WP200, where comparisons of flux rates are also carried out.

7 Intercomparison results for Band Imager (BI)

This section summarizes the intercomparison results for the BI instruments. All details are reported in ANNEX C (see **Sect. 9**).

Comparisons of results obtained with the Kayrros and UPV algorithms as applied to Sentinel- 2 and Landsat-8/9 radiances have been carried out for extended time periods over target regions (see **Sect. 7.1**) and for single observations (see **Sect 7.2**).

7.1 Target region intercomparisons for extended periods

The continuous monitoring intercomparison focused on two target regions. One is in the Bugdayly gas field in Turkmenistan, where two sub-regions have been identified for detailed study referred to as TKM1 and TKM2. The second target is a compressor station in the Permian Basin located in the USA.

The intercomparison focused on three aspects: (i) Detection performance, (ii) Methane emission quantification and (iii) Analysis of missed plume detections. The findings can be summarized as follows:

7.1.1 Detection performance

A key element here is to understand the pros and cons of an automatic detection, as done by Kayrros, against a manual approach, as done by UPV. Key results for the 3 regions TKM1, TKM2 and USA for Sentinel-2 are shown in **Figure 7.1.1-1** and summarized in **Table 7.1.1-1** in terms of “Recall”, which is equal to the number of plumes detected divided by the total number of plumes.

As shown in **Figure 7.1.1-1**, UPV detected 55 Sentinel-2 plumes: 10 for TKM1, 38 for TKM2 and 7 for USA whereas Kayrros detected 43 Sentinel-2 plumes: 6 for TKM1, 28 for TKM2 and 9 for USA. This case-study is in a blind detection mode, meaning emitting sources are unknown. Consequently, we are lacking ground-truth data, which prevents us from identifying false positives. In a first approach, and after a manual review, we concluded that all plumes are true positives and the recall metric was calculated accordingly by region for the two providers.

As we can see in **Table 7.1.1-1**, there is a significant drop in recall for Kayrros in the two regions of Turkmenistan. For TKM2, given the high number of detected plumes, it seems that the infrastructure is a constant emitting source. Kayrros uses an automatic selection of images to build the background. By taking the closest images in the past, without manual check, it is possible images are used as background that contain a plume.

For additional results and corresponding Landsat 8/9 results please see Annex C.

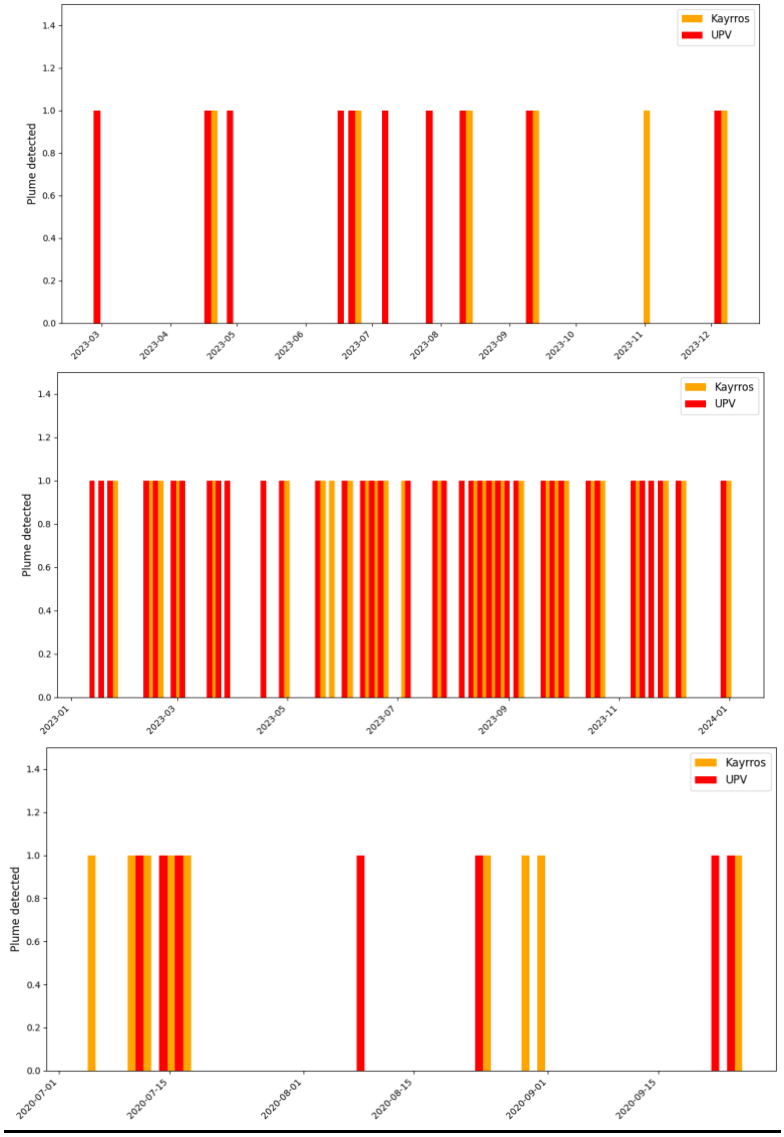


Figure 7.1.1-1: *Sentinel-2 detection comparisons between Kayrros (orange) and UPV (red), for TKM1 (upper), TKM2 (middle) and USA (lower).*

Region or sub-region	**Kayrros recall** "Automatic approach + manual labelling"	**UPV Recall** "Manual approach"
TKM1	0.55	0.91
TKM2	0.7	0.95
USA	0.82	0.64

Table 7.1.1-1: *Sentinel-2 detection performance by region for Kayrros and UPV. The recall is equal to the number of plumes detected divided by the total number of plumes.*

7.1.2 Methane emission quantification

A significant number of plumes (36) have been identified by both UPV and Kayrros for Sentinel-2. **Figure 7.1.2-1** presents point charts for each region comparing quantifications from both data providers. The emission estimates show systematic differences (UPV higher than Kayrros) but still overlap within their uncertainties. Orthogonal Distance Regressions were performed region by region to evaluate the statistical validity of this observation. Two “outliers” were excluded at this stage due to their excessive influence on the regression: 20231205 TKM1 and 20231230 TKM2. For TKM1 and USA, regression coefficients indicate UPV quantifications are approximately 64% - 73% higher than Kayrros on average. While R^2 values are relatively good for these regions, this must be considered alongside the limited sample sizes. Notably, the regression coefficients for TKM1, TKM2 and USA are relatively similar (1.64 vs 1.78 vs 1.73).

Since both Kayrros and UPV employ the IME methodology for plume emission rate quantification, the systematic bias in the IME formula can originate from either Ueff or IME/L parameters. **Figure 7.1.2-2** demonstrates that Ueff differences are usually minimal between the two methodologies. This outcome is anticipated since both UPV and Kayrros utilize identical wind data sources (ERA-5 Reanalysis) and apply the same effective wind calculation formula. The only methodological distinction lies in temporal handling: Kayrros uses the nearest past date for wind data, while UPV employs time interpolation. However, this interpolation approach can generate significant discrepancies, reaching up to 38% (as observed for TKM1 on June 23, 2023).

Figure 7.1.2-3 shows that IME/L is indeed responsible for the bias between Kayrros and UPV emission rate estimates. As the two parameters IME and L are positively correlated, it is hard to distinguish one of the two.

Figure 7.1.2-4 shows that there can be a significant difference in the plume delineation: Kayrros mask is bigger, more detailed than the one by UPV. To properly assess the impact of the plume mask, it is necessary to use the same L2 product for the quantification. This topic has been explored in a more detailed analysis for several single observations (see **Sect 7.2**).

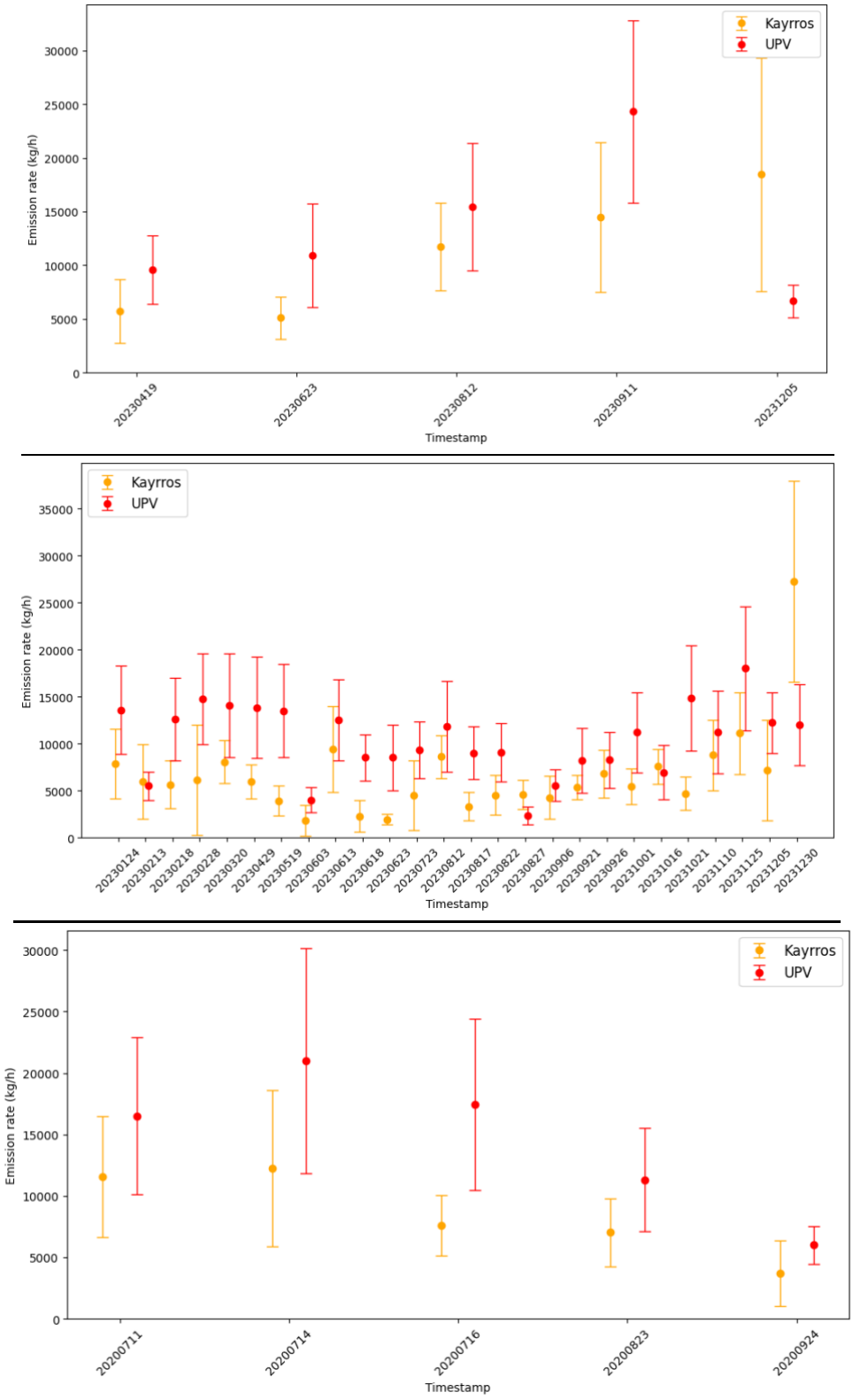


Figure 7.1.2-1: *Emission rate comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes for TKM1 (upper), TKM2 (middle) and USA (lower).*

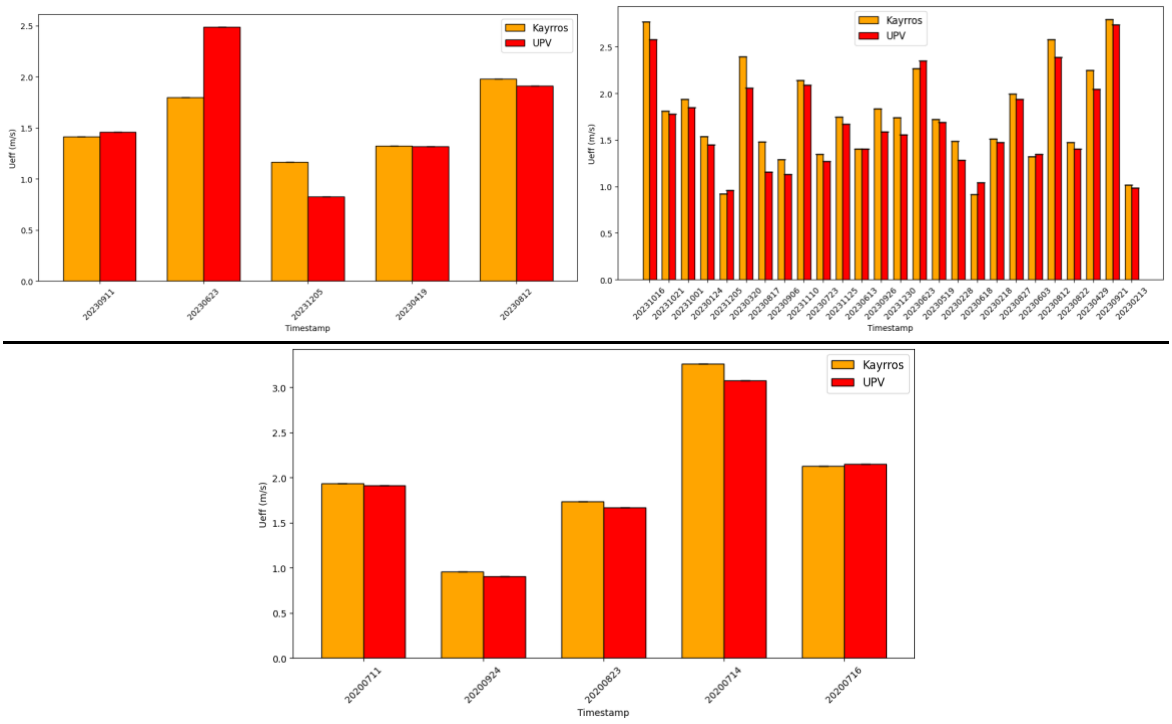


Figure 7.1.2-2: *Effective wind (Ueff) comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes, for TKM1 (upper left), TKM2 (upper right) and USA (lower).*

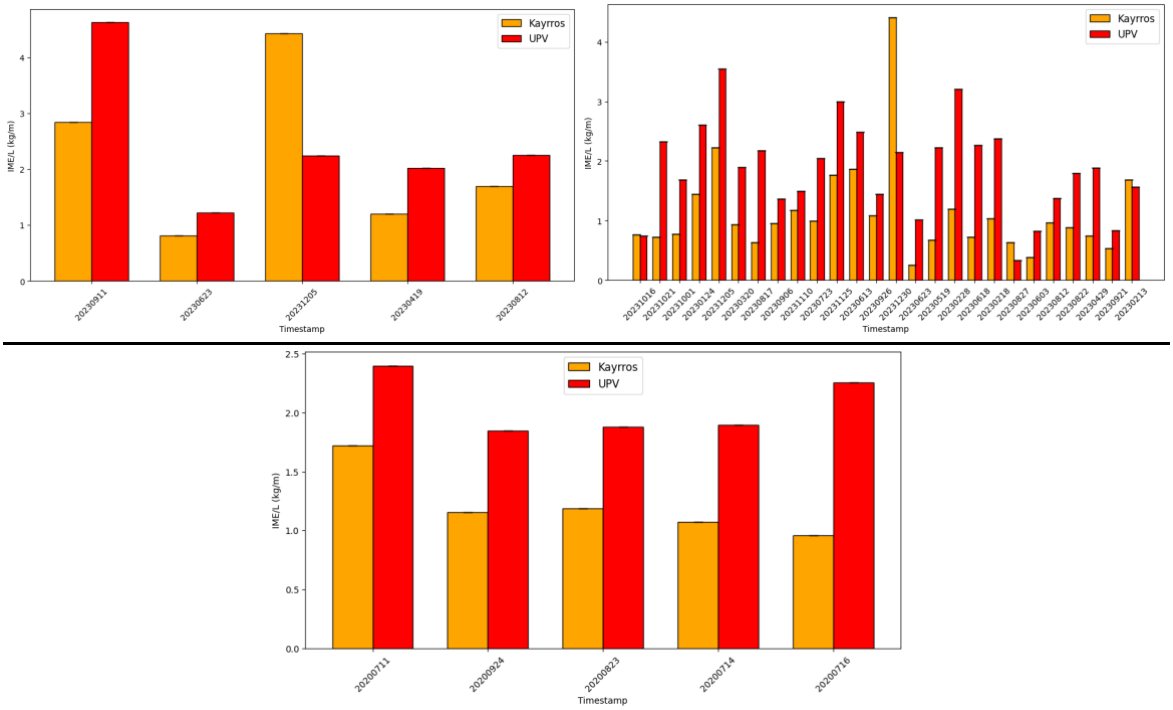


Figure 7.1.2-3: *IME/L comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes, for TKM1 (upper left), TKM2 (upper right) and USA (lower).*

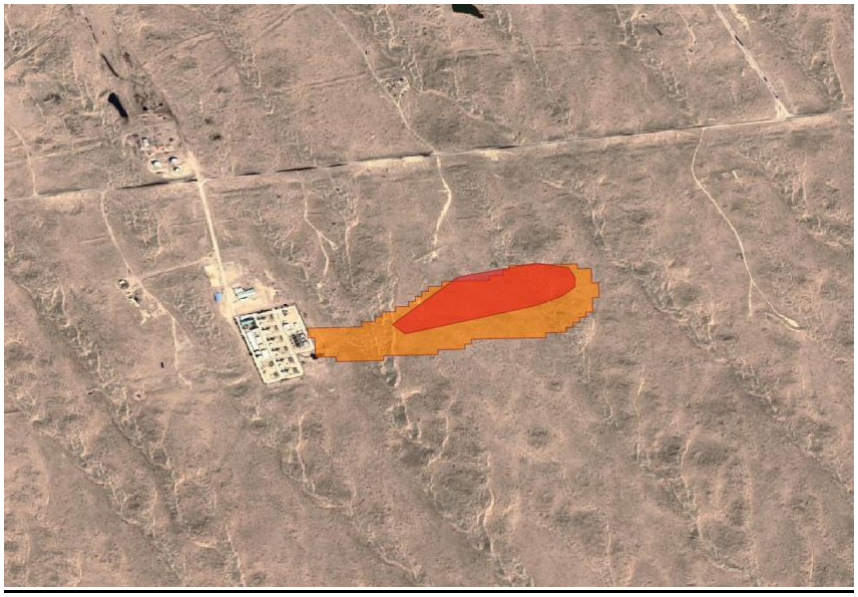


Figure 7.1.2-4: *Kayrros (orange) and UPV (red) masks delineated for 20230623 Sentinel-2 plume in TKM1.*

For additional results and corresponding Landsat 8/9 results (resulting in similar findings) please see Annex C.

7.1.3 Analysis of missed plume detections

This task aimed at identifying the root causes behind plumes detected by one provider but missed by the other. Understanding these factors is crucial for potential improvements. In deep learning applications for specific tasks like "methane plume detection," performance degradation typically stems from training set biases.

For Sentinel-2, no clear pattern emerges to explain plumes missed by either provider. For Landsat-8/9, Kayrros' neural network appears to regularly miss lower-intensity plumes (characterized by lower IME/L values). For detailed results please see Annex C.

7.2 Single observation intercomparisons

Two scenes have been identified for detailed intercomparison for Sentinel-2 and Landsat: An oil and gas field in Iraq and an oil and gas field in Venezuela. The Iraq scene (***Figure 7.2-1 left***) is a desert scene with a complex heterogeneous background, e.g., due to the soot of many producing flaring stacks. The chosen Venezuela site (***Figure 7.2-1 right***) is characterized by dense vegetation with seasonal variations, frequent cloud coverage and a lot of flaring. All these elements are likely to hinder the background removal step, which is essential in band imagers methane retrieval algorithms.

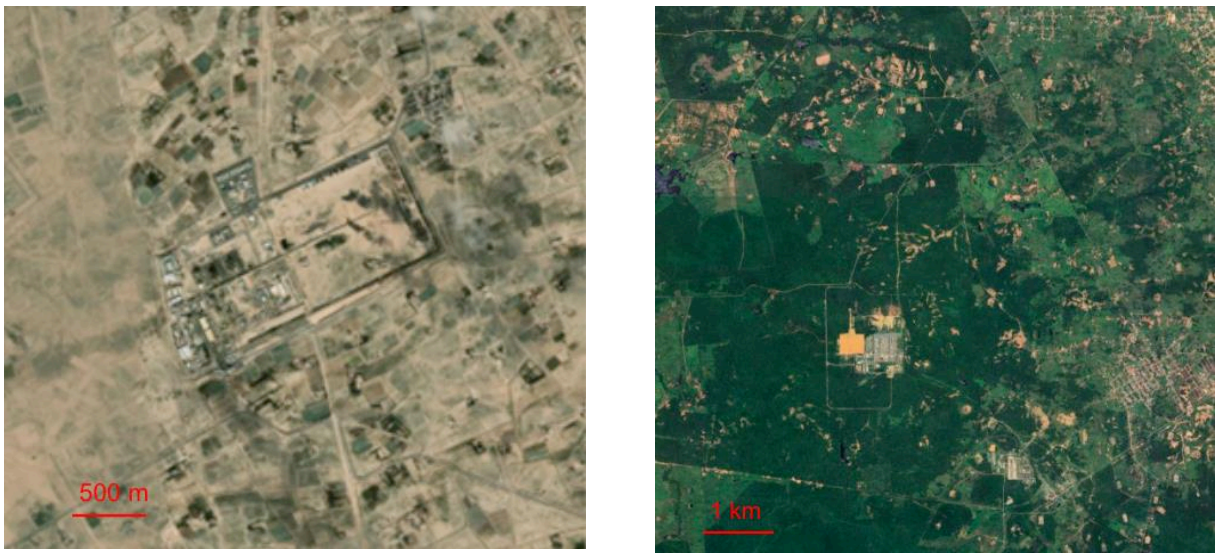


Figure 7.2-1: *O&G field in Iraq (left) (location: 47.742532, 30.274139) and O&G field in Venezuela (right) (location: -64.532524, 9.390828).*

The challenging observational conditions of the two scenes were anticipated to result in significant discrepancies between the two retrieval methods. The L2 products from both providers were all very noisy, with high values of methane enhancement due to the heterogenous background. Over the 8 scenes, only one plume was detected and it was by UPV, on the Sentinel-2 Iraq scene acquired on 2022, January 25th. Kayrros did not report any plume for this scene. Visual analysis of Figure 7.2-2 confirms the absence of discernible plume patterns or strong enhancements in the Kayrros L2 product at the location of the UPV delineated plume.

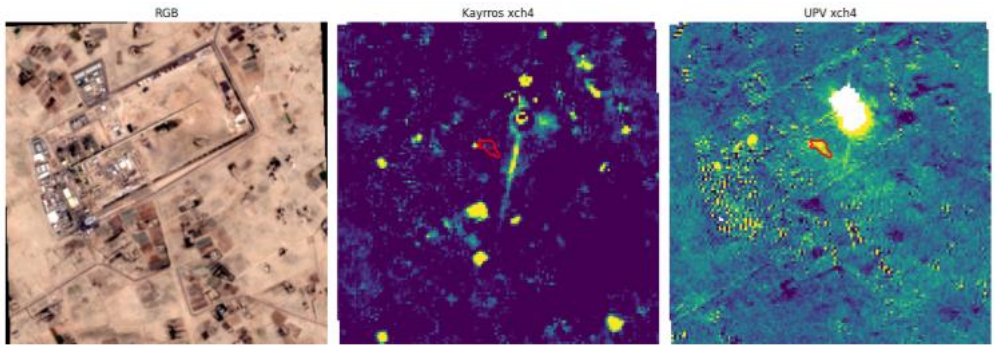


***Figure 7.2-2**: Comparison of L2 products (in ppm) between Kayrros (middle) and UPV (right) for the Sentinel-2 Iraq scene on 25/01/2022. The color scales differ between the Kayrros and UPV L2 products. The UPV delineation is displayed in red on the two images. The left image corresponds to the RGB composite image.*

In-depth comparison of the UPV and Kayrros L2 products reveals several differences in retrieval characteristics:

Cloud handling: Clouds present significant challenges for both algorithms. In Kayrros retrievals, clouds are not masked but exhibit values at or near zero. UPV applies cloud masking: however, imperfect mask boundaries result in artificially intense values at cloud edges in several scenes.

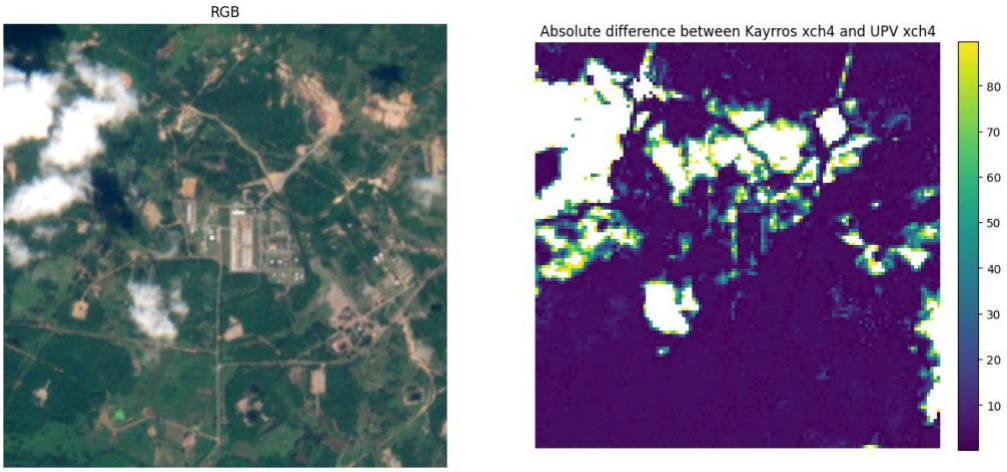


***Figure 7.2-3**: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Venezuela scene on 22/11/2023. The left image corresponds to the RGB composite image.*

Flare and smoke artefacts: Gas flares produce contrasting signatures between the two methods. Kayrros retrievals show flares as high-intensity features, while UPV retrievals exhibit suppressed flare signals.

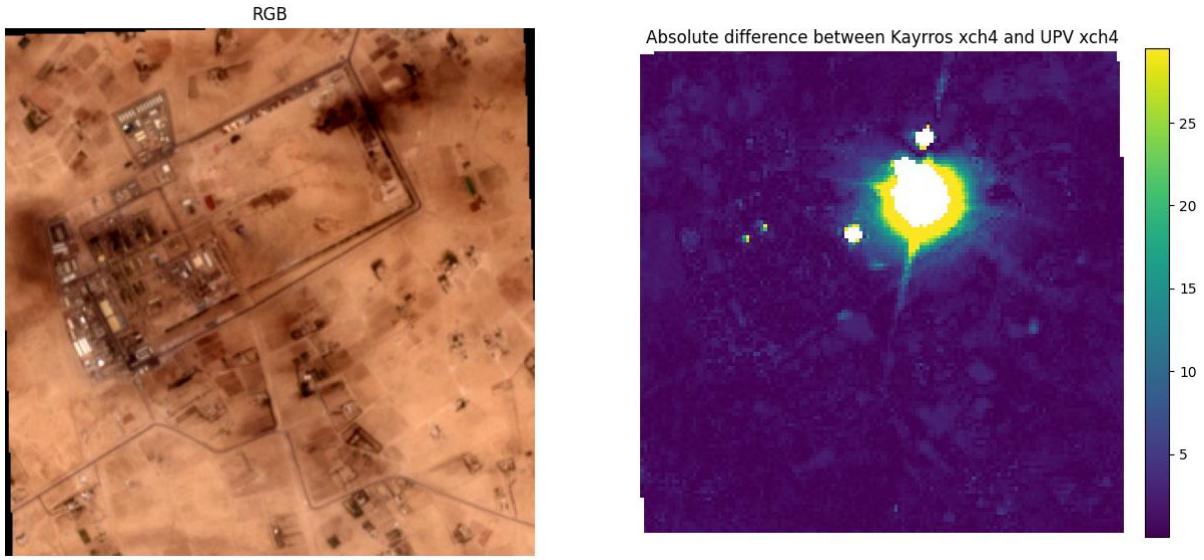


***Figure 7.2-4**: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Iraq scene on 22/11/2023. The left image corresponds to the RGB composite image.*

Urban areas: Spatial regions containing buildings typically exhibit more noise in both retrieval products.

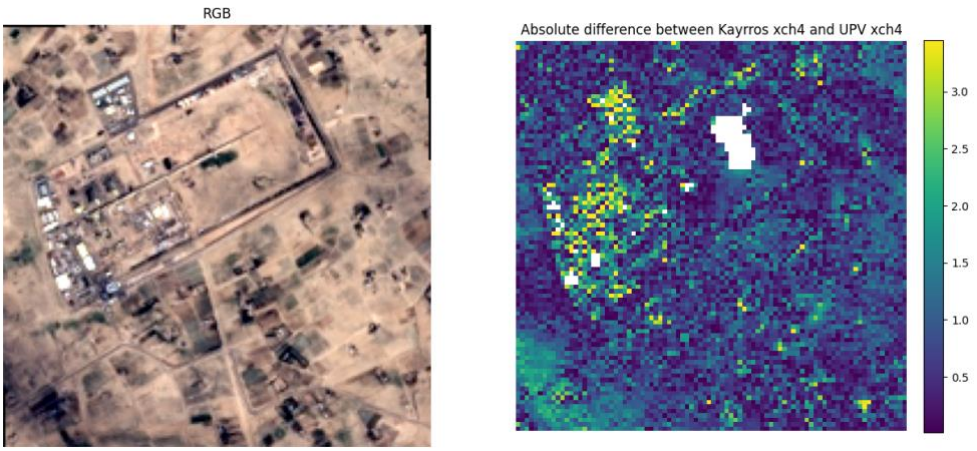


***Figure 7.2-5**: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Iraq scene on 20/12/2022. The left image corresponds to the RGB composite image.*

Overall, the L2 products demonstrate two primary differences: Kayrros retrievals exhibit higher noise levels with lower methane enhancement magnitudes compared

to UPV products. This increased noise is likely attributable to the automated background selection methodology, where automatic selection of ten historical reference images may introduce unexpected artifacts into the retrieval process. This observation is corroborated by quantitative analysis. The mean absolute error between Kayrros and UPV L2 products exhibit substantial magnitude, particularly pronounced for elevated XCH4 concentrations.

For GOES, the intercomparison is done on the two case studies presented in Groshenry et al. 2024:

- An emission in Mexico (-104.3308, 26.0941) on 12th May 2019.
- An emission in Idaho (-116.6542, 43.7443) on 12th October 2023.

For GOES, **Figure 7.2-6** demonstrates that for these Mexico and Idaho events, the relative difference between UPV and Kayrros estimates is approximately 40%, consistent with the magnitude of the differences observed for Sentinel-2 and Landsat-8/9. The key distinction lies in the non-overlapping uncertainty bars, attributed to the low relative uncertainties of both providers: 5% and 6% for UPV vs 3% and 6% for Kayrros.

To understand these differences detailed assessments have been carried out and all corresponding results are shown in Annex C. Here only one example is shown: **Figure 7.2-7** shows selected results from a detailed examination of one timestamp for the Idaho case: the comparison reveals significant masking discrepancies and very different noise levels.

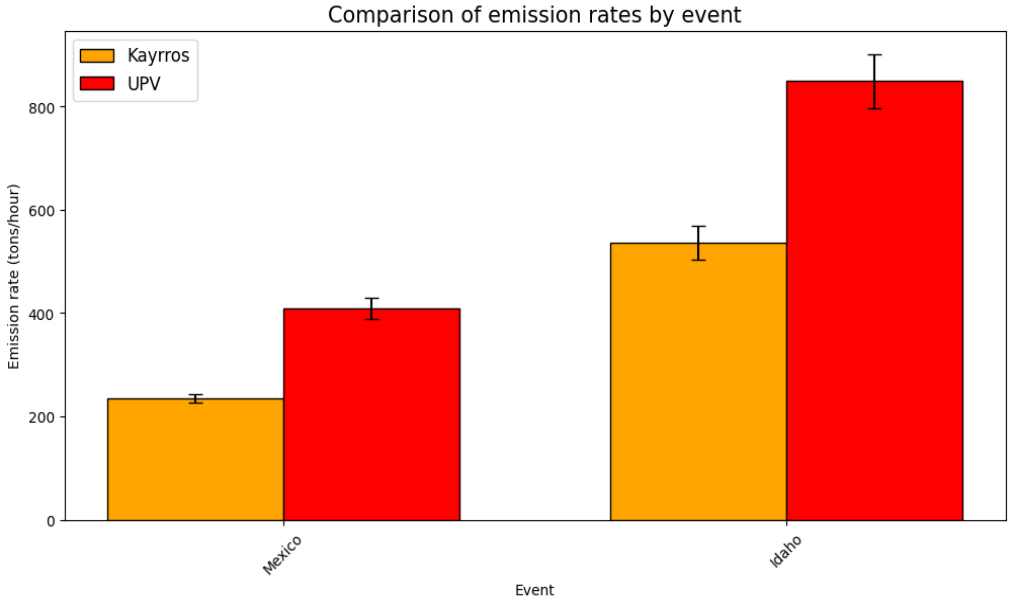


Figure 7.2-6: *Comparison of emission rates for GOES events.*

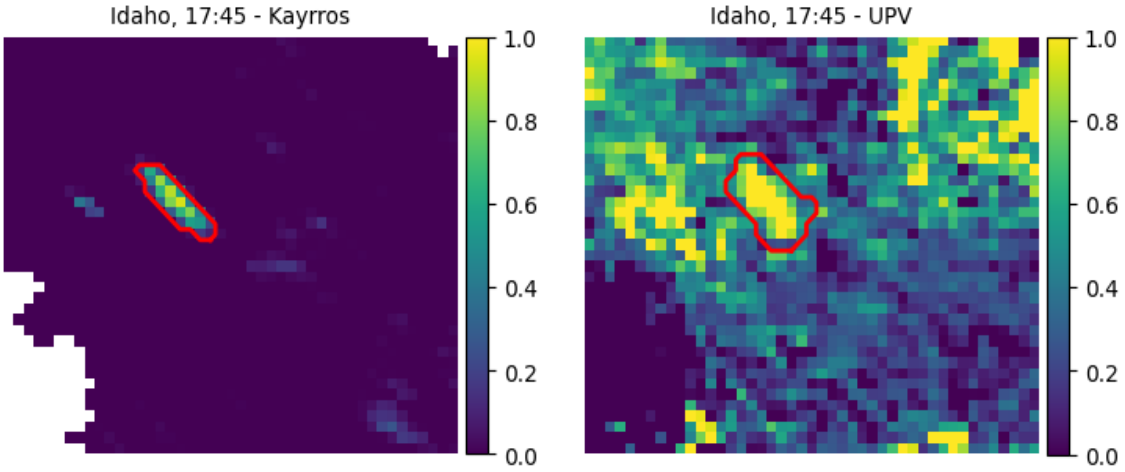


Figure 7.2-7: *Comparison between Kayrros and UPV retrievals and masks at 17:45 UTC for the Idaho event. On the left, Kayrros methane retrieval (in ppm) with the Kayrros mask delineated in red. On the right, UPV methane retrieval (in ppm) with the UPV mask delineated in red.*

The analysis further reveals substantial discrepancies between retrieval methods, clearly identifying methane retrieval algorithms as the primary source of flow rate differences between UPV and Kayrros quantifications. **Figure 7.2-8** demonstrates that methane retrieval distributions within the mask intersection areas exhibit significant statistical differences. Notably, the first quartile of UPV quantification values exceeds the third quartile of Kayrros values in both cases, indicating systematic bias between the two retrieval methodologies.

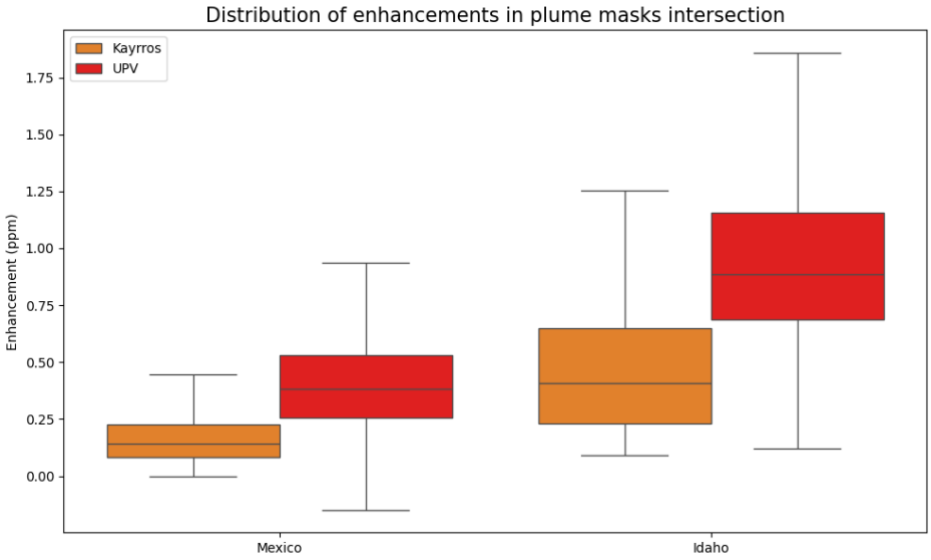


Figure 7.2-8: *Boxplots of methane retrieval in the intersection of the two masks, for each of the two events.*

Retrieval methodology differences are evident in background pixel estimation, where Kayrros exhibits predominantly zero-valued pixels. While Kayrros demonstrates lower plume intensity compared to UPV, it produces fewer artifacts and reduced high-value anomalies outside designated mask areas. It is also noteworthy that cloud pixels flagged as nodata in the Kayrros retrieval correspond to regions where UPV retrievals exhibit relatively high methane concentration values. **Figure 7.2-9** clearly illustrates these major differences in the two retrievals.

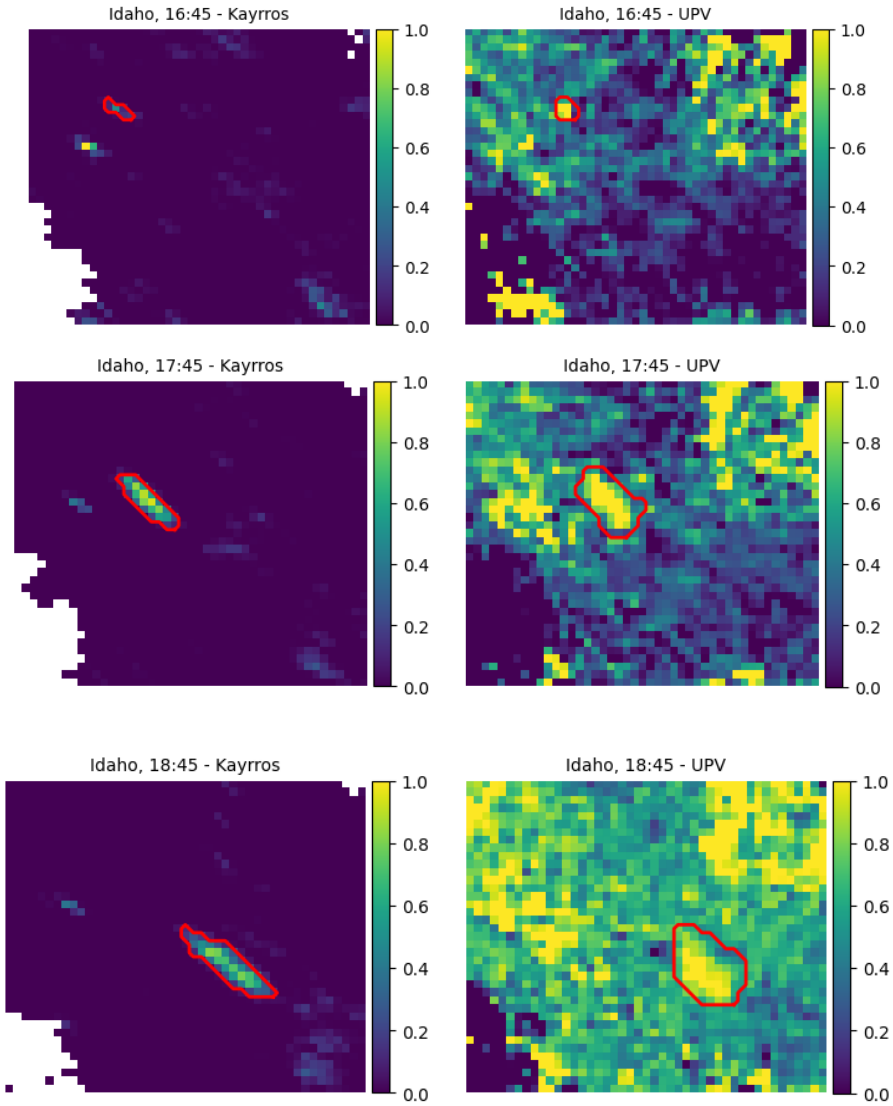


Figure 7.2-9: *Comparison of methane retrieval at three different timestamps for the Idaho event. On the left, Kayrros methane retrieval (in ppm) with the Kayrros mask delineated in red. On the right, UPV methane retrieval (in ppm) with the UPV mask delineated in red.*

7.3 Potential future research

The following topics have been identified:

- Across all band imagers, a significant difference was observed in the methane concentration maps generated by UPV and Kayrros. Identifying the primary cause of these differences is a crucial next step. Future research will investigate whether the disparity comes from the background removal or the methane retrieval algorithm itself.

- For Sentinel-2 and Landsat-8/9, based on the target region intercomparison conclusions, it would be interesting to deep dive into the intermediary results of the Kayrros deep learning detection model to understand the significant number of missed plumes by Kayrros. The aim would be to understand if the model completely missed the plume or if it detected a plume-like pattern but assigned a confidence score below the pre-defined confidence threshold. The single image intercomparison should be replicated on less challenging conditions, such as those encountered in the two Turkmenistan areas used for the target region intercomparison.

Concerning GOES, a similar comparative analysis could be conducted on a broader set of plumes, to significantly increase the statistical robustness of the conclusions. Furthermore, the current uncertainty computation, which is only derived from the uncertainty of the linear regression fitted to the Integrated Mass Enhancement (IME) time-series during the emission rate, could be more robust by including the uncertainty derived from the methane concentration retrieval itself. On this final topic, we could benefit from the conclusions of MEDUSA WP400.

8 References

Bösch, H., and M. Hilker, Inventory of point source emissions of CH4 estimated from high-resolution satellite data, Technical Report EU HORIZON-CL5-2022-D1-02 project EYE-CLIMA, deliverable 1.4, version 1.0, pp. 33, 31.12.2024, available from https://eyeclima.eu/products/public-reports/, 2024.

Buchwitz, M., Roger, J., Groshenry, A., Schneising-Weigel, O., Hilker, M., Noël, S., Reuter, M., Hachmeister, J., Boesch, H., Dogniaux, M., Zhang, X., Maasakkers, J. D., Aben, I., Gorroño, J., Guanter, L., Algorithm Intercomparison Plan (AIP), WP310 D3.1, version 1.1, Doc ID: SRON-ESG-PL-2024-003, Date: 16-12-2024, pp. 59, 2024.

Chambolle, A.: An Algorithm for Total Variation Minimization and Applications. Journal of Mathematical Imaging and Vision 20, 89–97, https://doi.org/10.1023/B:JMIV.0000011325.36760.1e, 2004.

Carvalho, D.: An Assessment of NASA's GMAO MERRA-2 Reanalysis Surface Winds, J. Climate, 32, 8261–8281, https://doi.org/10.1175/JCLI-D-19-0199.1, 2019.

Duren, R., Thorpe, A., Foster, K., Rafiq, T., Hopkins, F., Yadav, V., Bue, B., Thompson, D., Conley, S., Colombi, N., Frankenberg, C., McCubbin, I., Eastwood, M., Falk, M., Herner, J., Croes, B., Green, R., and Miller, C.: California's methane super-emitters, Nature, 575, 180–184, https://doi.org/10.1038/s41586-019-1720-3, 2019.

Ehret T., De Truchis A., Mazzolini M., Morel J-M., and Facciolo G., "Automatic methane plume quantification using sentinel-2 time series," in IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1955–1958, 2022.

Ember DUH, Brief: Urgency to update Germany's coal mine methane emission factor, https://ember-climate.org/insights/in-brief/de-undermines-cmm-emissions/, Ember Climate and Deutsche Umwelthilfe (DUH), including TROPOMI analysis supporting material and press release (https://www.duh.de/fileadmin/user_upload/download/Pressemitteilungen/Energie/Methan/Embargoed_Public_of_Ember_Brief__Urgency_to_update_methane_emission_factors_from_Germany's_coal_mining_sector.pdf), 10-Apr-2024, 2024.

Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., and Conley, S.: Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region, P. Natl. Acad. Sci. USA, 113, 9734–9739, https://doi.org/10.1073/pnas.1605617113, 2016.

Fuentes Andrade, B., Buchwitz, M., Reuter, M., Bovensmann, H., Richter, A., Boesch, H., and Burrows, J. P.: A method for estimating localized CO_2 emissions from co-located satellite XCO_2 and NO_2 images, Atmos. Meas. Tech., 17, 1145–1173, https://doi.org/10.5194/amt-17-1145-2024, 2024.

Gloudemans, A. M. S., Schrijver, H., Hasekamp, O. P., and Aben, I.: Error analysis for CO and CH_4 total column retrievals from SCIAMACHY 2.3 µm spectra, Atmos. Chem. Phys., 8, 3999–4017, https://doi.org/10.5194/acp-8-3999-2008, 2008.

Groshenry, A., Giron, C., Lauvaux, T., d'Aspremont, A. & Ehret, T. Detecting methane plumes using prisma: deep learning model and data augmentation. In NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning (NIPS, 2022). https://www.climatechange.ai/papers/neurips2022/17, 2022

Groshenry A., Giron C., Hessel C., de Franchis C., Facciolo G., Ehret T., "Methane emissions monitoring using geostationary satellites" in IGARSS 2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024.

Guanter, L., I. Irakulis-Loitxate, J. Gorroño, E. Sánchez-García, D. H. Cusworth, D. J. Varon, S. Cogliati, and R. Colombo, "Mapping methane point emissions with the prisma spaceborne imaging spectrometer," Remote Sensing of Environment, vol. 265, p. 112671, 2021.

Guanter, Luis, Javier Roger, Shubham Sharma, Adriana Valverde, Itziar Irakulis-Loitxate, Javier Gorroño, Xin Zhang, Berend J. Schuit, Joannes D. Maasakkers, Ilse Aben, Alexis Groshenry, Antoine Benoit, Quentin Peyle and Daniel Zavala-Araiza, Multisatellite Data Depicts a Record-Breaking Methane Leak from a Well Blowout, 30 June 2024, Environmental Science & Technology Letters, DOI: 10.1021/acs.estlett.4c00399, 2024.

Heidinger et al., ALGORITHM THEORETICAL BASIS DOCUMENT ABI Cloud Mask, https://www.star.nesdis.noaa.gov/goesr/documents/ATBDs/Baseline/ATBD_GOES-R_Cloud_Mask_v3.0_Jul2012.pdf

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, ERA Report Series, https://doi.org/10.21957/tkic6g3wm, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloy- aux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. Roy. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Jacob, D. J., Turner, A. J., Maasakkers, J. D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., and Frankenberg, C.: Satellite observations of atmospheric methane and their value for quantifying methane emissions, Atmos. Chem. Phys., 16, 14371–14396, https://doi.org/10.5194/acp-16-14371-2016, 2016.

Jacob, D. J., Varon, D. J., Cusworth, D. H., et al., Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane, Atmos. Chem. Phys., 22, 9617–9630, doi:10.5194/acp-22-9617-2022, 2022.

Krautwurst, S., Gerilowski, K., Borchardt, J., Wildmann, N., Gałkowski, M., Swolkień, J., Marshall, J., Fiehn, A., Roiger, A., Ruhtz, T., Gerbig, C., Necki, J., Burrows, J. P., Fix, A., and Bovensmann, H.: Quantification of CH_4 coal mining emissions in Upper Silesia by passive airborne remote sensing observations with the Methane Airborne MAPper (MAMAP) instrument during the CO2 and Methane (CoMet) campaign, Atmos. Chem. Phys., 21, 17345–17371, https://doi.org/10.5194/acp-21-17345-2021, 2021.

Lauvaux, T., Giron, C., Mazzolini, M., d'Aspremont, A., Duren, R., Cusworth, D., Shindell, D. and Ciais, P.,. Global assessment of oil and gas methane ultra-emitters. Science, 375(6580), pp.557-561, 2022.

Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D., and Saltelli, A.: Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy, Reliab. Eng. Syst. Safe., 206, 107300, https://doi.org/10.1016/j.ress.2020.107300, 2021.

Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., aan de Brugh, J., Schneider, A., Wu, L., Hase, F., Kivi, R., Wunch, D., Pollard, D. F., Shiomi, K., Deutscher, N. M., Velazco, V. A., Roehl, C. M., Wennberg, P. O., Warneke, T., and Landgraf, J.: Methane retrieved from TROPOMI: improvement of the data product and validation of the first 2 years of measurements, Atmos. Meas. Tech., 14, 665–684, https://doi.org/10.5194/amt-14-665-2021, 2021.

Lorente, A., Borsdorff, T., Martinez-Velarte, M. C., and Landgraf, J.: Accounting for surface reflectance spectral features in TROPOMI methane retrievals, Atmos. Meas. Tech., 16, 1597–1608, https://doi.org/10.5194/amt-16-1597-2023, 2023.

Maasakkers, J. D., Varon, D. J., Elfarsdóttir, A., McKeever, J., Jervis, D., Mahapatra, G., Pandey, S., Lorente, A., Borsdorff, T., Foorthuis. L. R., Schuit, B. J.,Tol, P., van Kempen, T. A., van Hees, R., Aben, I., Using satellites to uncover large methane emissions from landfills, Sci. Adv., Vol 8, Issue 32, DOI: 10.1126/sciadv.abn9683, 2022.

Molod, A., Takacs, L., Suarez, M., Bacmeister, J., Song, I.-S., and Eichmann, A.: The GEOS-5 Atmospheric General Circu- lation Model: Mean Climate and Development from MERRA to Fortuna, Technical Report Series on Global Modeling and Data Assimilation, Tech. Rep., https://ntrs.nasa.gov/citations/ 20120011790 (last access: 25 July 2023), 2012.

Naus, S., J. D. Maasakkers, R. Gautam, M. Omara, R. Stikker, A. K. Veenstra, B. Nathan, I. Irakulis-Loitxate, L. Guanter, S. Pandey, M. Girard, A. Lorente, T. Borsdorff, and I. Aben, Assessing the Relative Importance of Satellite-Detected Methane Superemitters in Quantifying Total Emissions for Oil and Gas Production Areas in Algeria, Environmental Science & Technology, 57 (48), 19545-19556, DOI: 10.1021/acs.est.3c04746, 2023.

Pei, Zhipeng, Ge Han, Huiqin Mao, Cuihong Chen, Tianqi Shi, Keyi Yang, Xin Ma, Wei Gong, Improving quantification of methane point source emissions from imaging spectroscopy, Remote Sensing of Environment, Volume 295, 113652, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2023.113652, 2023.

Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., Bovensmann, H., and Burrows, J. P.: Towards monitoring localized CO2 emissions from space: co-located regional CO_2 and NO_2 enhancements observed by the OCO-2 and S5P satellites, Atmos. Chem. Phys., https://www.atmos-chem-phys.net/19/9371/2019/, 19, 9371-9383, 2019.

Roger, J., Irakulis-Loitxate, I., Valverde, A., Gorrono, J., Chabrillat, S., Brell, M., Guanter, L.: High-Resolution Methane Mapping With the EnMAP Satellite Imaging Spectroscopy Mission. - IEEE Transactions on Geoscience and Remote Sensing, 62, 1-12, https://doi.org/10.1109/TGRS.2024.3352403, 2024.

Sadavarte, P., Pandey, S., Maasakkers, J. D., Lorente, A., Borsdorff, T., Denier van der Gon, H., Houweling, S., and Aben, I.: Methane Emissions from Superemitting Coal Mines in Australia Quantified Using TROPOMI Satellite Observations, Environ. Sci. Technol., 55, 16573–16580, https://doi.org/10.1021/acs.est.1c03976, 2021.

Sherwin, E. D., Rutherford, J. S., Chen, Y., Aminfard, S., Kort, E. A., Jackson, R. B., and Brandt, A. R.: Single-blind validation of space-based point-source methane emissions detection and quantification, Scientific Reports, 13, 3836, https://doi.org/10.1038/s41598-023-30761-2, 2023.

Sherwin, E. D., El Abbadi, S. H., Burdeau, P. M., Zhang, Z., Chen, Z., Rutherford, J. S., Chen, Y., and Brandt, A. R.: Single-blind test of nine methane-sensing satellite systems from three continents, Atmos. Meas. Tech., 17, 765–782, https://doi.org/10.5194/amt-17-765-2024, 2024.

Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., and Burrows, J. P.: Severe Californian wildfires in November 2018 observed from space: the carbon monoxide perspective, Atmos. Chem. Phys., 20, 3317-3332, https://doi.org/10.5194/acp-20-3317-2020, 2020a.

Schneising, O., Buchwitz, M., Reuter, M., Vanselow, S., Bovensmann, H., and Burrows, J. P.: Remote sensing of methane leakage from natural gas and petroleum systems revisited, Atmos. Chem. Phys., 20, 9169-9182, https://doi.org/10.5194/acp-20-9169-2020, 2020b.

Schneising, O., Buchwitz, M., Hachmeister, J., Vanselow, S., Reuter, M., Buschmann, M., Bovensmann, H., and Burrows, J. P.: Advances in retrieving XCH_4 and XCO from Sentinel-5 Precursor: improvements in the scientific TROPOMI/WFMD

algorithm, Atmos. Meas. Tech., 16, 669–694, https://doi.org/10.5194/amt-16-669-2023, 2023.

Schneising, O., Buchwitz, M., Reuter, M., Weimer, M., Bovensmann, H., Burrows, J. P., and Bösch, H.: Towards a sector-specific CO/CO_2 emission ratio: Satellite-based observation of CO release from steel production in Germany, Atmos. Chem. Phys., 24, 7609–7621, https://doi.org/10.5194/acp-24-7609-2024, 2024.

Schuit, B. J., Maasakkers, J. D., Bijl, P., Mahapatra, G., van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D. J., McKeever, J., Jervis, D., Girard, M., Irakulis-Loitxate, I., Gorroño, J., Guanter, L., Cusworth, D. H., and Aben, I.: Automated detection and monitoring of methane super-emitters using satellite data, Atmos. Chem. Phys., 23, 9071–9098, https://doi.org/10.5194/acp-23-9071-2023, 2023.

Stein, A.F., Draxler, R.R, Rolph, G.D., Stunder, B.J.B., Cohen, M.D., and Ngan, F., (2015). NOAA's HYSPLIT atmospheric transport and dispersion modeling system, Bull. Amer. Meteor. Soc., 96, 2059-2077, http://dx.doi.org/10.1175/BAMS-D-14-00110.1, 2015.

Thompson, D. R., Thorpe, A. K., Frankenberg, C., Green, R. O., Duren, R., Guanter, L., Hollstein, A., Middleton, E., Ong, L., and Ungar, S.: Space-based remote imaging spectroscopy of the Aliso Canyon CH4 superemitter, Geophys. Res. Lett., 43, 6571–6578, https://doi.org/10.1002/2016GL069079, 2016.

Thorpe, A. K., Frankenberg, C., and Roberts, D. A.: Retrieval techniques for airborne imaging of methane concentrations using high spatial and moderate spectral resolution: application to AVIRIS, Atmos. Meas. Tech., 7, 491–506, https://doi.org/10.5194/amt-7-491-2014, 2014.

Tu, Q., Hase, F., Schneider, M., García, O., Blumenstock, T., Borsdorff, T., Frey, M., Khosrawi, F., Lorente, A., Alberti, C., Bustos, J. J., Butz, A., Carreño, V., Cuevas, E., Curcoll, R., Diekmann, C. J., Dubravica, D., Ertl, B., Estruch, C., León-Luis, S. F., Marrero, C., Morgui, J.-A., Ramos, R., Scharun, C., Schneider, C., Sepúlveda, E., Toledano, C., and Torres, C.: Quantification of CH_4 emissions from waste disposal sites near the city of Madrid using ground- and space-based observations of COCCON, TROPOMI and IASI, Atmos. Chem. Phys., 22, 295–317, https://doi.org/10.5194/acp-22-295-2022, 2022.

Tu, Q., Hase, F., Qin, K., Cohen, J. B., Khosrawi, F., Zou, X., Schneider, M., and Lu, F.: Quantifying CH_4 emissions from coal mine aggregation areas in Shanxi, China, using TROPOMI observations and the wind-assigned anomaly method, Atmos. Chem. Phys., 24, 4875–4894, https://doi.org/10.5194/acp-24-4875-2024, 2024.

Vanselow, S., Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., Boesch, H., and Burrows, J. P.: Automated detection of regions with persistently enhanced methane concentrations using Sentinel-5 Precursor satellite data, Atmos. Chem. Phys., 24, 10441–10473, https://doi.org/10.5194/acp-24-10441-2024, 2024.

Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., et al., Satellite discovery of anomalously large methane point sources from oil/gas

production. Geophysical Research Letters, 46, 13,507–13,516. https://doi.org/10.1029/2019GL083798, 2019.

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, Atmos. Meas. Tech., 11, 5673–5686, https://doi.org/10.5194/amt-11-5673-2018, 2018.

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 Sens. Environ., 120, 70–83, https://doi.org/10.1016/j.rse.2011.09.027, 2012.

Worden, J., Green, P., Eldering, A. and Sherwin, E., Common Practices for Quantifying Methane Emissions from Plumes Detected by Remote Sensing, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.IR.8575, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=960016 (Accessed July 3, 2025), 2025.

Zhang, Y., Gautam, R., Pandey, S., Omara, M., Maasakkers, J. D., Sadavarte, P., Lyon, D., Nesser, H., Sulprizio, M. P., Varon, D. J., Zhang, R., Houweling, S., Zavala-Araiza, D., Alvarez, R. A., Lorente, A., Hamburg, S. P., Aben, I., and Jacob, D. J.: Quantifying methane emissions from the largest oil-producing basin in the United States from space, Science Advances, 6, eaaz5120, https://doi.org/10.1126/sciadv.aaz5120, 2020.

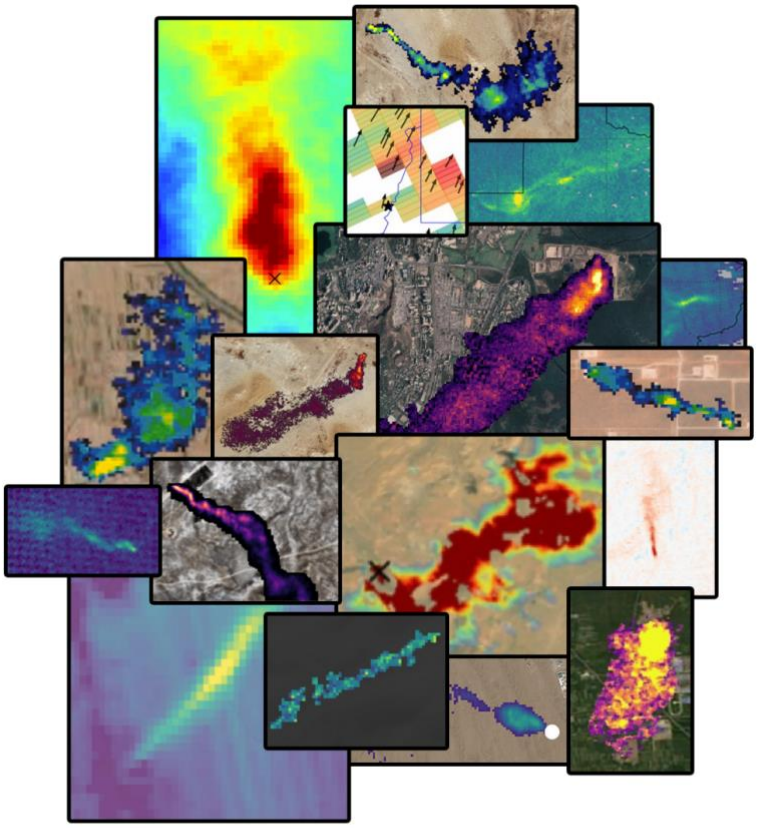
9 ANNEXes A-C

The following pages – if attached – contain the following 3 documents:

- ANNEX A: Flux Mapper (FM) comparison results
- ANNEX B: Hyperspectral Imager (HI) comparison results
- ANNEX C: Band Imager (BI) comparison results

ANNEX A (FM comparisons) to Algorithm Intercomparison Report (AIR)

WP320 – Annex for Deliverable D3.2



Authors

Lead author:

- Michael Buchwitz: AIR document lead and FM (Flux Mapper) sections lead (incl. ANNEX A for FM)

All authors:

Institute of Environmental Physics, University of Bremen, Germany (IUP-UB)
- Michael Buchwitz
- Stefan Noël
- Jonas Hachmeister
- Oliver Schneising-Weigel
- Michael Hilker
- Maximilian Reuter

SRON Netherlands Institute for Space Research, Leiden, The Netherlands
- Matthieu Dogniaux
- Xin Zhang
- Joannes D. Maasakkers
- Ilse Aben

Kayrros, Paris, France
- Julian Akani-Guéry
- Quentin Peyle

Change log

Version	Date	Status	Authors	Reason for change
Draft 1	5-June-2025	Submitted for internal use	(*)	New document
Draft 2	24-June-2025	Internal draft	(*)	Additional assessments related to target regions added
Draft 3	27-August-2025	Internal draft submitted for internal use	(*)	First full draft with inputs from SRON and Kayrros added + additional IUP-UB results
0.9	1-October-2025	Submitted to team for final check / inputs	(*)	Team inputs added.
0.9.1	13-October-2025	Submitted to I. Aben (IA) for review	(*)	Only harmonization of version number and date with other AIR documents; no change of content
0.9.2	10-December-2025	Submitted	(*)	Improved based on minor comments and corrections from IA
1.0	18-February-2026	Accepted	-	Version number change after acceptance

(*) see Authors list

Table of Contents

1. Executive summary

This document is an ANNEX to the Algorithm Intercomparison Report (AIR) of ESA project MEDUSA (Methane Emissions Detection Using Satellites Assessment).

The objective of MEDUSA is to develop the techniques needed for a pre-operational system able to harmonize and integrate global information on subnational (e.g., urban areas) to facility scale anthropogenic methane emissions derived using diverse satellite instruments and algorithms.

This document is ANNEX A to AIR focusing on intercomparison results for Flux Mapper (FM). Within MEDUSA algorithm intercomparisons are grouped into three groups depending on satellite instrument type. These three instrument types are:

- Flux Mapper (FM): These are moderate spatial resolution (several kilometers) sensors. At this stage of the project this a single sensor, namely TROPOMI on Sentinel-5 Precursor (S5P).
- Hyperspectral Imager (HI): PRISMA, EnMAP and EMIT/ISS. Intercomparison results for HI are presented separately in ANNEX B.
- Band Imager (BI): Sentinel-2 (S2), Landsat-8 (L8), Landsat-9 (L9) and GOES satellites. Intercomparison results for BI are presented separately in ANNEX C.

The algorithm intercomparisons have been carried out as described in the MEDUSA Algorithm Intercomparison Plan (AIP). As described in the AIP, the FM intercomparisons focus on the intercomparison of methane emission estimates as obtained by the application of three different algorithms to XCH_4 Level 2 products from TROPOMI/S5P. These methane emission estimation algorithms are:

- The Integrated Mass Enhancement (IME) algorithm used by SRON.
- The inversion algorithm used by Kayrros.
- The Cross-Sectional-Flux (CSF) algorithm used by IUP-UB.

The intercomparison focusses on year 2021 and on two complementary types of assessments: Results from all three algorithms have been compared in detail for 10 pre-defined target regions (as defined in the AIP) and in addition a comparison of the global data as obtained using the SRON and Kayrros algorithms has been carried out. The IUP-UB algorithm currently uses as input the location (in terms of latitude and longitude) of the target of interest. In contrast, the SRON and Kayrros algorithms can be applied to all data as the identification of the location of emission sources (or more precisely: the identified origin of the detected/observed emission plume) is part of the algorithm, i.e., approximate target locations are an output of the algorithms of SRON and Kayrros.

Note that the algorithms used here (i.e., for document AIR) are only a sub-set of the algorithms used in the MEDUSA project. Included here are only those methods where more than one methane emission inversion algorithm is available (within this project) to obtain hotspot emission information for a given satellite sensor. In a parallel task,

called “Systematic product intercomparison and validation”, results from additional sensors and algorithms are used and assessed.

The intercomparison results presented in detail in this document can be summarized as follows:

Target region intercomparisons:

Detailed comparisons of methane emission estimates (Level 4 products) at 10 target locations have been presented in this section as obtained from TROPOMI satellite XCH_4 retrievals (Level 2 products) using the 3 different Level 2 to Level 4 algorithms as developed and used by SRON, Kayrros and IUP-UB. The comparison shows that, as expected, the IUP-UB Cross-Sectional-Flux (CSF) algorithm provides typically the largest number of methane emission estimates for a given target location. This is because this algorithm is not based on emission plume detection in contrast to the other 2 algorithms. Instead, only a small set of selection criteria is used in order to decide if the satellite data can be used for emission estimation or not: Data coverage must be dense enough (e.g., > 90% of the area used for the CSF analysis needs to be covered by “good” (e.g., cloud free) data independent of the spatial pattern of the methane concentration field) and wind speed should not be too low (> 1 m/s). If these criteria are met than the quality flag (QF) attached to the corresponding emission estimate is set to “good” (QF=0) (otherwise it is set to 1). Consequently, estimated emissions using IUP-UB’s CSF algorithm may also be zero or even negative in case no methane plume is present and because XCH_4 retrievals are noisy. The smallest number of emission estimates are provided by the Kayrros algorithm due to a conservative detection method and their quantification approach, which is based on detailed modelling, which is challenging due to the required good knowledge of meteorological information (e.g., wind field), issues related to turbulence, etc. The number of estimates provided by SRON is in between the number of estimates of Kayrros and IUP-UB.

It needs to be pointed out that the IUP-UB CSF algorithm as used for MEDUSA is new and still under development. Most of the results generated by IUP-UB for MEDUSA have been produced with version 1.6 of this algorithm. CSF v1.6 is also the algorithm used for emission estimates from the Hyperspectral Imagers (HI) EnMAP, PRISMA and EMIT (see Annex B to document AIR). Results reported in the following refer to comparisons with v1.6 except if stated otherwise.

Primarily due to the low number of Kayrros estimates and the fact that often only one algorithm provided an estimate for a given target location and satellite overpass, the number of collocations is small. At 3 of the 10 target locations no collocations have been found. For the 7 other locations where collocations exist, on average good to reasonable agreement between the IUP-UB and the SRON estimates has been found for 3 of these targets but for the 4 other targets, the IUP-UB estimates were significantly higher (on average by more than a factor of 2). In contrast, the agreement between Kayrros and SRON was found to be much better with agreement typically within some 10%. But this finding is based on only a very small number of collocations and more meaningful are the global comparisons results reported separately.

Several assessments have been carried out to explain the relative overestimation of the IUP-UB CSF v1.6 emission estimates compared to SRON at some of the target locations. It has been identified that the different Level 2 input data are not the driver for these differences, and it is argued that also the wind speed used by IUP-UB is unlikely the reason for a factor of 2 or larger differences. IUP-UB uses for TROPOMI the Planetary Boundary Layer (PBL) average of the wind speed, which is in line with Varon et al., 2019, who found that the PBL average wind speed is appropriate for TROPOMI CSF-based emission estimates.

One must note that the target region comparisons are based on only a small number of targets and with only a small number of collocations per target. This raises the question how representative the target region comparisons are. Additional comparisons for several other locations have been conducted independently by B. Dils (BIRA-IASB) in the context of MEDUSA work package (WP) 220. These comparisons show on average very good agreement between the emission estimates as obtained by IUP-UB with CSF v1.6, SRON and Kayrros. Here the main findings based on correlation (scatter) plots are: slope $\beta = 1.01$ ($N = 26$, $R^2 = 0.74$) for SRON vs IUP-UB, $\beta = 0.92$ ($N = 22$, $R^2 = 0.69$) for Kayrros vs IUP-UB and $\beta = 1.05$ ($N = 22$, $R^2 = 0.55$) for SRON vs Kayrros. This indicates that all three algorithms result in very similar emission estimates on average (as all slopes are close to 1) although (as indicated by the moderate R^2 values) the individual emission estimates may vary significantly (by often several 10%). This indicates that the level of agreement or disagreement of the results obtained with the different algorithms depends on which subset is compared.

Nevertheless, additional sensitivity studies have been carried out to find out if the identified differences of the IUP-UB CSF v1.6 results at the target regions relative to SRON can be explained. These activities resulted in a new version 1.7 of the IUP-UB CSF algorithm. The main differences of v1.7 compared to v1.6 are an improved background correction and a waiver of the 3-hour plume length limit used for v1.6. It is shown that v1.7 results in much better agreement with the SRON estimates at the selected targets indicating that the implemented changes to the IUP-UB CSF algorithm are related to aspects responsible for at least large parts of the observed differences.

So far only a limited amount of data has been processed and analyzed with v1.7 and further studies are needed to find out if v1.7 is better than v1.6 in general, i.e., also for cases not yet investigated.

Due to the limited number of collocations with the Kayrros estimates at the selected 10 target regions, the target region comparison focused on comparisons between the IUP-UB and SRON algorithms. For a more meaningful comparison of the SRON and the Kayrros algorithms, please see the following summarizing the global comparisons.

Global intercomparisons:

Global datasets of methane emission plumes detected in TROPOMI data have been reported by Kayrros and SRON for the year 2021. Both datasets have detections covering the main methane emitting regions (e.g Permian basin, West coast of Turkmenistan, Shanxi coal mines, etc.), with SRON also covering additional locations (e.g. Casablanca, Istanbul or coal mines in Poland) consistent with a three-times higher number of detections (3233 SRON versus 1060 Kayrros). Plumes from both datasets have been matched based on the comparison of their reported plume masks. The Kayrros – SRON matching plumes (N=643, with 306 that include emission quantifications in both) agree on average (small mean difference) and show a good correlation (r = 0.62). This result is quite remarkable considering that Kayrros and SRON quantification methods are strictly different (except for the input TROPOMI data). However, the scatter of Kayrros – SRON differences is quite significant, with a 53% relative difference standard deviation. Considering the 30% and 40% mean relative uncertainties reported by Kayrros and SRON (resp.), 77% of the 306 plumes agree within uncertainty range. Most of the plume-to-plume emission difference scatter is shown to be explained by the difference between Kayrros and SRON emission quantification methods: using SRON's quantification method on Kayrros detected plumes reduces difference variance by 83%. This comparison of Kayrros – SRON matching plumes also brought up useful improvement points for both algorithms: SRON's effective wind calibration are slightly inconsistent between 10 m and PBL-averaged wind speeds; Kayrros' handling of missing TROPOMI pixels appears to worsen comparison results. Examining the 386 plumes detected by Kayrros but not reported by SRON, we assess that filtering (39% due to TROPOMI pixel filters and 35% due to human labelling of detections) employed by the SRON TROPOMI plume detection framework can explain the large majority of these non-detected or non-reported plumes. Conversely, of the 2568 SRON-only detections, 70% correspond to short and low-intensity plumes, which Kayrros' anomaly-based computer vision algorithm tends to miss due to its reliance on smoothed input data. Additional differences (15%) stem from borderline cases, where variations in pipeline design and input data can cause detections to fall into different filtering categories across teams.

2. Overview satellite sensors

This section provides a short overview of the satellite sensors relevant for the MEDUSA algorithm intercomparison. These are sensors where more than one methane emission inversion algorithm is used and evaluated within MEDUSA. This is a sub-set of the satellite sensors used within MEDUSA.

This document focusses on intercomparison results for Flux Mapper (FM). Within MEDUSA algorithm intercomparisons are grouped into three groups depending on satellite instrument type. These three instrument types are:

- Flux Mapper (FM): These are moderate spatial resolution (several kilometer) sensors. At this stage of the project this a single sensor, namely TROPOMI on Sentinel-5 Precursor (S5P).
- Hyperspectral Imager (HI): PRISMA, EnMAP and EMIT/ISS. Intercomparison results for HI are presented separately in ANNEX B.
- Band Imager (BI): Sentinel-2 (S2), Landsat-8 (L8), Landsat-9 (L9) and GOES satellites. Intercomparison results for BI are presented separately in ANNEX C.

2.1. Flux Mapper (FM)

2.1.1. TROPOMI on Sentinel-5 Precursor (S5P)

TROPOMI is the only payload instrument of the ESA Sentinel-5 Precursor mission launched in October 2017 (Veefkind et al., 2012). It is a push-broom imaging spectrometer measuring radiances in eight spectral bands from the ultraviolet (UV) to the shortwave infrared (SWIR), allowing the retrieval of various atmospheric constituents including methane.

TROPOMI has the unique ability to combine observations at city-scale spatial resolution with global coverage on a daily basis due to its large 2600 km swath consisting of individual measurements with a footprint size of 5.5 x 7 km^2 at nadir (7x7 km^2 before 6 August 2019) in the SWIR bands (7 and 8) relevant for methane retrieval.

Input for this study are Level 2 (L2) XCH_4 data products from TROPOMI/S5P, where XCH_4 is the column-averaged dry-air mole fraction of atmospheric methane. Used are two L2 products: (i) the operational data product, which is based on the retrieval code developed by SRON (e.g., Lorente et al., 2021, 2023) and (ii) the scientific data product generated by University of Bremen based on the WFMD (Weighting Function Modified DOAS) retrieval algorithm (e.g., Schneising et al., 2023).

3. Overview methane plume detection & emission rate estimation algorithms

This section provides a brief overview of methane plume detection and emission rate estimation algorithms as used and assessed within the MEDUSA algorithm intercomparison task.

For a general overview on methods and community-accepted practices for quantifying methane emissions based on plumes detected via spectroscopic remote sensing please see Worden et al., 2025, from which **Fig. 3-1** was taken, or Jacob et al., 2022. Here we use results obtained with the Integrated Mass Enhancement (IME) method and the Cross-Sectional Flux (CSF) method (and an inversion method based on detailed modelling not listed in that figure).

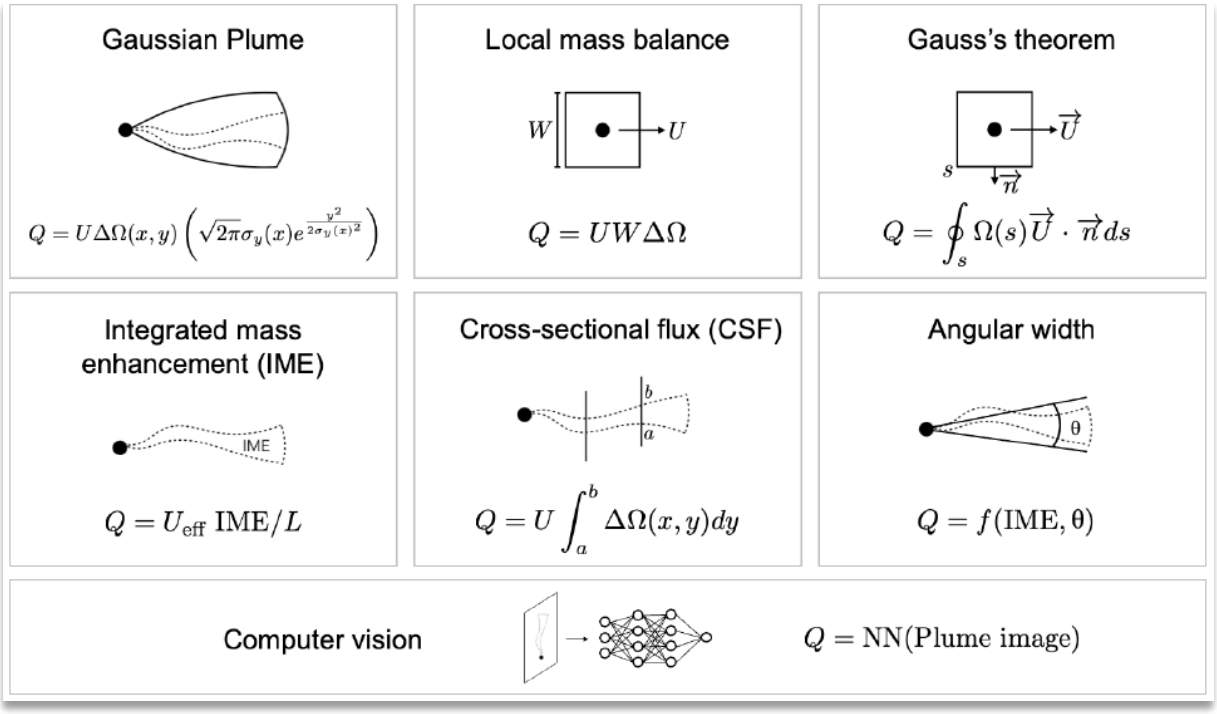


Figure 3-1: *Overview of methods used to obtain methane emission estimates (Q) from atmospheric methane vertical columns (Ω) or column enhancements (ΔΩ) and wind speed (U). Figure from Worden et al., 2025 (based on Jacob et al., 2022).*

In this AIR document algorithms to obtain methane emission information are also referred to as Level 4 (L4) algorithms. These algorithms use as input lower-level data products such as Level 2 (atmospheric methane information from and for individual satellite footprints or ground pixel) or Level 1 (i.e., radiance) products. For FM (i.e., TROPOMI/S5P) Level 2 XCH_4 products are used as input for this project.

3.1. Algorithms for Flux Mapper (FM)

Three different Level 2 to Level 4 algorithms are described in this section, which are used to generate methane emission estimates (Level 4 products) from the TROPOMI/S5P XCH_4 retrievals (Level 2 products). These methane emission estimation algorithms are:

- The Integrated Mass Enhancement (IME) algorithm used by SRON.
- The inversion algorithm from Kayrros.
- The Cross-Sectional-Flux (CSF) algorithm developed by IUP-UB.

3.1.1. SRON algorithm for TROPOMI/S5P

The base SRON TROPOMI methane plume estimation algorithm is based on the Integrated Mass Enhancement (IME) method, implemented to quantify plumes automatically detected in TROPOMI data through a 2-step machine learning procedure. Each detection is independently verified by two human operators. This scheme is extensively described in Schuit et al., 2023. TROPOMI input data for the SRON methane emission estimation algorithm is the ESA operational Level 2 (L2) XCH_4 data product derived from TROPOMI radiance spectra (Level 1 (L1) product) through the SRON L1 to L2 retrieval algorithm.

To detect plumes, each orbit of TROPOMI data is swept by a 32 x 32 pixel moving window approach with 50% overlap in both across- and along-track directions. Every 32 x 32 pixel image is first analyzed by a Convolutional Neural Network (CNN) determining if the image contains a plume-like signal in the methane field and, if positive, subsequently examined by a Support Vector Classifier (SVC) that further analyzes the detection for possible artefacts (e.g. due to albedo features).

In case of positive detection, the Class Activation Map (CAM) of the CNN is multiplied by the methane enhancement values (calculated by subtracting a local background) contained in the 32 x 32 pixel image to identify the pixel with the highest enhancement and confidence as belonging to the plume. Starting from this pixel, a plume mask is generated using an iterative outward dilation approach, including pixels with enhancements 1.8 standard deviations or more above the image mean in the plume mask. The plume mask is final once the pixel selection converges and is used to calculate the methane emission rate associated with this plume. The most upwind pixel in the plume mask is identified as the most likely source location. This location is hence purely determined based on TROPOMI data and is not cross-referenced against possible emitters that can be seen in visual imagery of the area.

The quantification method relies on the widely used IME method, first introduced by Frankenberg et al., 2016, for aircraft measurements, and first applied to satellite data by Varon et al., 2018. It relates the plume emission rate Q with the excess of methane contained in the plume mask denoted as IME (computed with a background value

corresponding to the median methane column in the 32 x 32 pixel image outside the plume mask), with the plume length L defined as the square-root of the mask area and with the effective wind speed U_{eff}, calibrated on atmospheric transport simulations against the 10m wind speed U_{10m} or the wind speed averaged over the Planetary Boundary Layer thickness U_{PBL}. We have $Q = IME \times U_{eff} / L$; with $U_{eff} = 0.47 \times U_{PBL} + 0.31$, or $U_{eff} = 0.59 \times U_{10m}$.

The IME method is used in an ensemble approach to yield both an average emission rate and an uncertainty estimate. A 43923-member ensemble is built by varying different inputs to the IME quantification method:

- The wind speed product:
  - ERA5 10m wind speed (Hersbach et al., 2020)
  - GEOS-FP 10m wind speed (Molod et al., 2012)
  - GEOS-FP (Molod et al., 2012) PBL wind speed
- The enhancement threshold for a pixel to be included in the plume mask
- The background concentration used to yield enhancement values
- The wind speed values
- The effective wind speed calibration coefficients

The last four are varied in steps. For example, the wind speed is perturbed by −50 % to +50 % in steps of 10 %. Finally, the average across the ensemble is reported as emission rate, and the standard deviation across the ensemble is reported as uncertainty. This approach is also described in more detail in MEDUSA WP410.

For specific cases requiring further attention, SRON can use the CSF quantification method and atmospheric inversions using atmospheric modeling (e.g., Maasakkers et al., 2019, and Guanter et al., 2024).

Plume detections including all supporting variables are provided in CSV files.

3.1.2. Kayrros algorithm for TROPOMI/S5P

Kayrros algorithm is based on the computation of a methane anomaly map, described in Lauvaux et al., 2021, applied to the TROPOMI/S5P operational L2 data V02.05.00.

Continuous groups of positive pixels in the anomaly map are selected as plume candidates, and contiguous (but distinct) plumes are separated using watershed segmentation.

The detections are systematically validated by two human expert labelers who filter out false positives. They do this by verifying the presence of a potential emission source and assessing the consistency of the detection based on geometric features, wind alignment, and albedo.

Each detection is quantified using a corresponding simulated plume, calculated with the Lagrangian particle dispersion model HYSPLIT (Stein et al., 2015). The simulation is conducted in forward mode on a 0.01x0.01-degree grid, using meteorological data

from the Global Forecast System (GFS) provided by the National Centers for Environmental Prediction (NCEP). The emission rate is finally obtained by normalizing the flow rate of the simulated plume with the ratio of enhancement between the actual observation and the simulation.

A flux rate uncertainty is estimated through a sensitivity analysis conducted on a batch of 200 plumes, evaluating the uncertainty arising from Sentinel-5P measurements, source location selection, background quantification, and meteorological data errors.

3.1.3. IUP-UB algorithm for TROPOMI/S5P

The IUP-UB methane emission estimation algorithm is based on the Cross-Sectional-Flux (CSF) approach. The CSF method is based on integrating the methane column enhancement (in, e.g., tons/km^2) times wind speed (in, e.g., km/hour) along transects perpendicular to wind direction to estimate the methane emission (in, e.g., tons/hour).

CSF algorithms have been developed and used at IUP-UB for TROPOMI/S5P methane (Schneising et al., 2020b) and carbon monoxide (Schneising et al., 2020a, 2024) emission estimation and to obtain CO_2 emission information from OCO-2 (Reuter et al., 2019) and OCO-3 (Fuentes Andrade et al., 2024). The IUP-UB TROPOMI/S5P methane L4 algorithm described in Schneising et al., 2020b, has been used to obtain emission estimates for large areas such as entire gas and oil fields (e.g., Schneising et al., 2020b) but also to estimates emissions of open pit coal mines (Borchardt et al., 2025).

For MEDUSA related applications a new algorithm has been implemented that can also be applied to other sensors such as EnMAP. The latest version of this CSF algorithm is version 1.6 and this version has been used for the results shown here.

Atmospheric methane input data from TROPOMI/S5P is the scientific XCH_4 Level 2 (L2) data product as generated with the Weighting Function Modified Differential Absorption Spectroscopy (WFM-DOAS or WFMD) retrieval algorithm (Schneising et al., 2023). We use the WFMD v1.8 L2 product as input for the results shown here but also show some results from the application of our CSF v1.6 method to the operational TROPOMI V02 L2 XCH_4 product.

The IUP-UB CSF algorithm has been applied to S5P XCH_4 retrievals around pre-defined target locations (see **Sect. 4.1**). The latitude and longitude of the target location of interest is input for this algorithm. A plume detection algorithm is not used but also not needed for this algorithm. Consequently, the obtained emission estimate can be zero or even negative (in case of no plume and a noisy methane image).

We rotate the original methane image around the target location such that the wind blows to the right (i.e., in the positive x-direction) and convert the XCH_4 (in ppb) into vertical column enhancements (in tons/km^2) (see **Figure 3.1.3-1**).

This rotation has the advantage that the cross-sections along which the methane column enhancements are integrated and oriented along the vertical (i.e., y-direction). Only cross-sections located in a box of size 100 km times 100 km are used and quantitatively analysed, where the left side of the box is centred at the target location. To fill data gaps and to obtain an estimate of the background concentration, also the surrounding area is used. Data gaps are filled using Gaussian smoothing. The background correction (used to obtain the desired methane enhancements) is based on the median plus a correction for contributions transported by wind into the box.

We compute emission estimates for several cross-sections for different distances from the target location. To minimize the impact of issues related with too short (e.g., complex plume shape close to the source) or too long distances from the target location, we only use emission estimates from cross-sections some small distance away from the target location and using only cross-sections not too far away from the source. For the latter we use a maximum distance corresponding to 3 hours after satellite overpass as computed using the wind speed. The reported emission estimate is the mean of the estimates as obtained for the valid cross-sections.

For wind information meteorological information from the European Centre for Medium-Range Weather Forecasts (ECMWF) the ERA5 reanalysis product (Hersbach et al., 2018) is used. We assume that on the several 10 km scale of interest here, the emitted methane is well-mixed in the Planetary Boundary Layer (PBL). We therefore use the PBL average of the wind (nearest neighbour to the target location) for the CSF method.

The main result of this CSF algorithm is an estimate of the methane emission including its (1-sigma) uncertainty for given target locations and for each satellite overpass, where enough data are available around the target of interest to obtain a robust emission estimate.

The unit of the emission estimate is tons/hour. The emission product files are in CSV format with each line corresponding to a single overpass. These lines contain (at least) the following information: target latitude and longitude, time, estimated emission and its (1-sigma) uncertainty (both in tons/hour), and the quality flag (0 = good).

The uncertainty is estimated by quadratically adding the following terms: (i) The standard deviation of the emission estimates as obtained for the individual cross-sections divided by the square root of the number of independent measurements (Nindep; computed using the method described in Fuentes Andrade et al., 2024), (ii) a wind related component considering wind variability and a general uncertainty of 0.5 m/s, and (iii) other contributors assuming an additional contribution of 20%.

The quality flag is set to 0, which means “good”, if (i) there are enough data in a box of size 100 x 100 km^2 located at the target location (> 90% of area covered; shown as “NdataFlag” in the figures, e.g., **Figure 3.1.3-1**), if (ii) Nindep larger than 0 (“NindepFlag”), and if (iii) wind speed is not too low (> 1 m/s; “WindFlag”).

Figure 3.1.3-1 shows as an example detailed CSF v1.6 results for a S5P overpass on 18-Sept-2021 of the Norte III landfill, Buenos Aires, Argentina.

Figure 3.1.3-2 shows the corresponding average map of the rotated methane enhancement and other information such as the estimated methane emissions from end of 2017 to end of 2023.

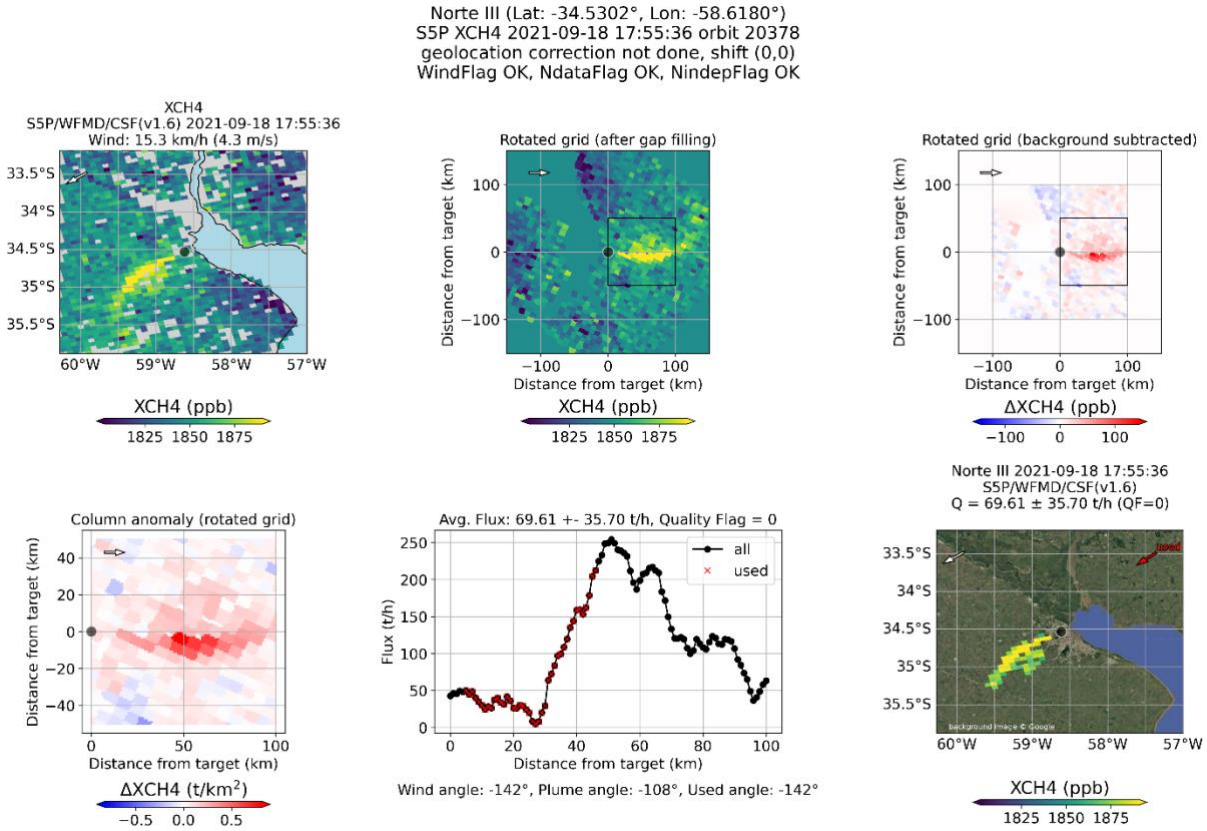


Figure 3.1.3-1: *IUP-UB CSF v1.6 methane emission estimation for the Norte III landfill, Buenos Aires, Argentina. Top left: XCH_4 map obtained by application of the WFMD v1.8 retrieval algorithm to TROPOMI/S5P radiances on 18-Sept-2021. Top middle: Rotated and gap filled methane map. Top right: Corresponding methane enhancement map with target location (black dot) and box (square shown via thin black lines) used for CSF computation. Bottom left: Column anomaly in box in tons/km^2. Bottom middle: Flux estimates for all (black symbols) and used (red symbols) individual cross-sections. Used cross-sections (red symbols) start some small distance away from the target and have a less than 3 hours time distance as estimated from wind speed. The title lists the estimated emission and its 1-sigma uncertainty and the value of the quality flag (0 = good). Bottom right: Identified methane enhancement near target (isolated plume not used for analysis, only for illustration) on top of background image from google maps.*

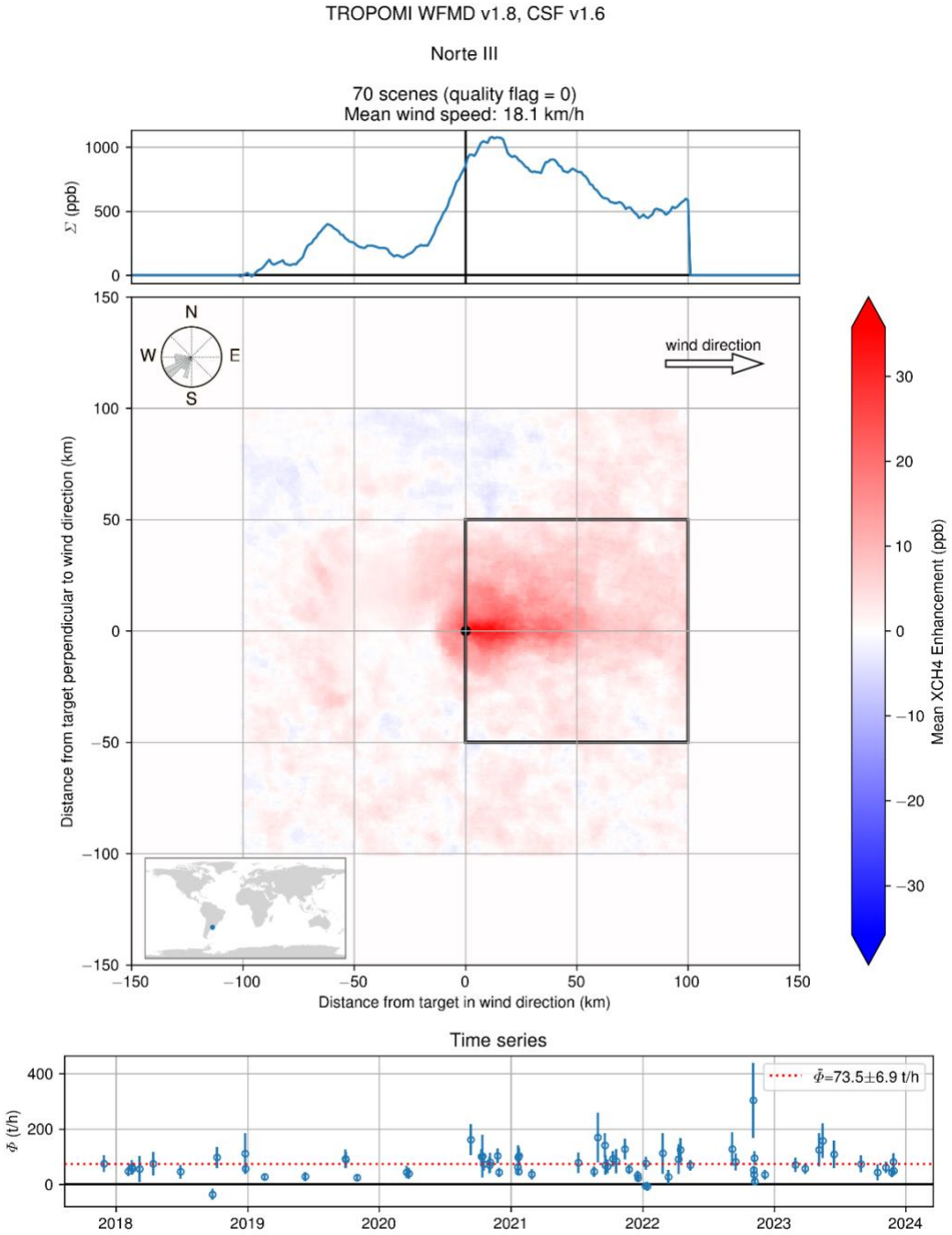


Figure 3.1.3-2: *Middle: Average methane enhancement map computed by averaging all individual rotated maps as obtained for the Norte III landfill target. Top: Corresponding mean methane enhancement as a function of distance from the source in wind direction. Bottom: Methane emission estimates for the individual overpasses.*

4. Intercomparisons for pre-defined target locations

In this section the intercomparison results for the 10 pre-defined target regions are presented. The global comparison results are shown in **Sect. 5**.

The target region intercomparison is based on the 3 data sets of estimated methane emissions as derived from TROPOMI/S5P using the algorithms from SRON, Kayrros, and IUP-UB (for details see **Sect. 3.1**).

4.1. Target locations

As described in document AIP (Buchwitz et al., 2024), 10 target regions have been defined. They are listed here in **Table 4.1-1**.

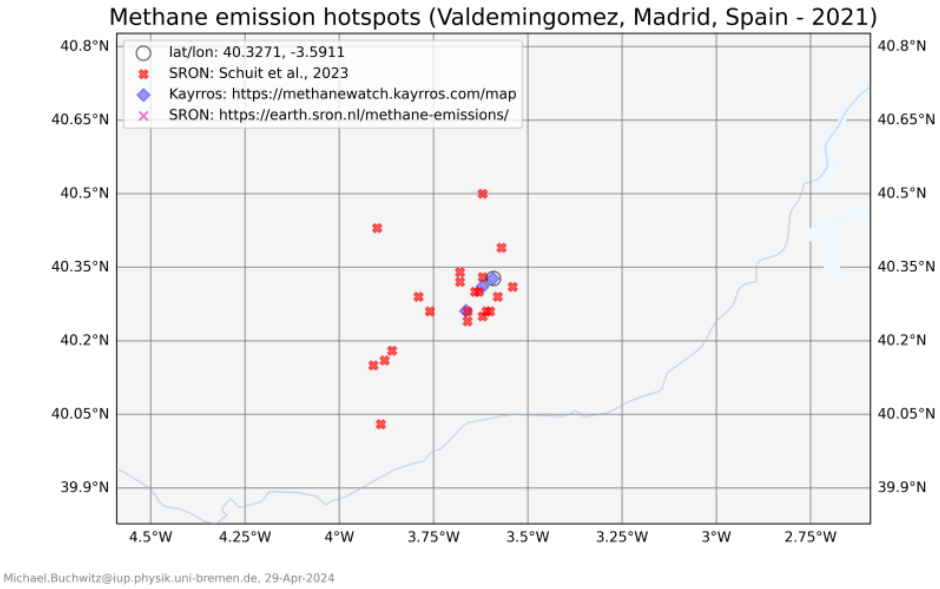
The target regions have been selected such that they cover different methane sources (coal, oil, gas, waste, …) and different observing conditions from good to challenging conditions. Good observation conditions typically correspond to high surface albedo, low solar zenith angles (high sun elevation), low cloud contamination, etc. Challenging observation conditions include observations near the coast (as water is a poor reflector in the SWIR spectral region) or areas with complex topography and inhomogeneous surface reflectivity.

Target regions are specified by location (latitude and longitude, see **Table 4.1-1**) but also the surrounding region needs to be considered as the locations of individual detections vary (e.g., because they are based on individual TROPOMI overpasses where plumes could be cutoff because of missing data). **Figure 4.1-1** illustrates this for TROPOMI/S5P methane plume origin detection. Therefore, all detections in a certain area around a given pre-defined location need to be included in the intercomparisons.

FM-No ID	Target short name	Target latitude (*)	Target longitude (*)	Comment
FM-01 **HAS**	Hassi_Messaoud	31.7500 N	6.1000 E	Algeria; oil and gas; (e.g., Naus et al., 2023)
FM-02 **KOR**	Korpezhe	38.6000 N	54.2500 E	Turkmenistan, west coast; oil and gas; (e.g., (Varon et al., 2019)
FM-03 **VAL**	Valdemingomez (Las Dehesas)	40.3271 N	3.5911 W	Spain, Madrid; landfill area (Tu et al., 2022; Krautwurst et al., 2024); major landfills within few kilometers: Las Dehesas (40.336427N, 3.590375W) and Pinto (40.257118N, 3.63755W)
FM-04 **NOR**	Norte_III	34.5302 S	58.6180 W	Argentina, Buenos Aires, landfill (Maasakkers et al., 2022)
FM-05 **USI**	Upper_Silesia_Zory	50.0000 N	18.6500 E	Poland; Upper Silesian Coal Basin (USCB) (Krautwurst et al., 2021)
FM-06 **HAI**	Hail_Creek	21.500 S	148.3800 E	Australia, Queensland, coal mining area (Sadavarte et al., 2021; Borchardt et al., 2025)
FM-07 **PER**	Permian_Pecos	31.3 N	103.2 W	USA, oil and gas fields (Schneising et al., 2020b; Zhang et al, 2020)
FM-08 **SHA**	Shanxi_Qinshui	35.6 N	112.55 E	China, coal mining area (Tu et al., 2024)
FM-09 **KAR**	Karaturun_East	45.3324 N	52.3730 E	Kazakhstan; oil field; 2023 (before, during and after blowout event); challenging as near the coast (Guanter et al., 2024)
FM-10 **HAM**	Hambach	50.9100 N	6.5300 E	Germany; open pit coal mining area (lignite); challenging; highlighted using S5P in Ember DUH, 2024

Table 4.1-1: *List of target regions for intercomparison of methane emissions derived from TROPOMI/S5P using the algorithms from SRON, Kayrros and IUP-UB. (*) Note that also the surrounding area (some 0.1 degrees latitude and longitude) is part of the target region as the identified locations vary to some extent (see* ***Fig. 4.1-1****).*

(a)



(b)

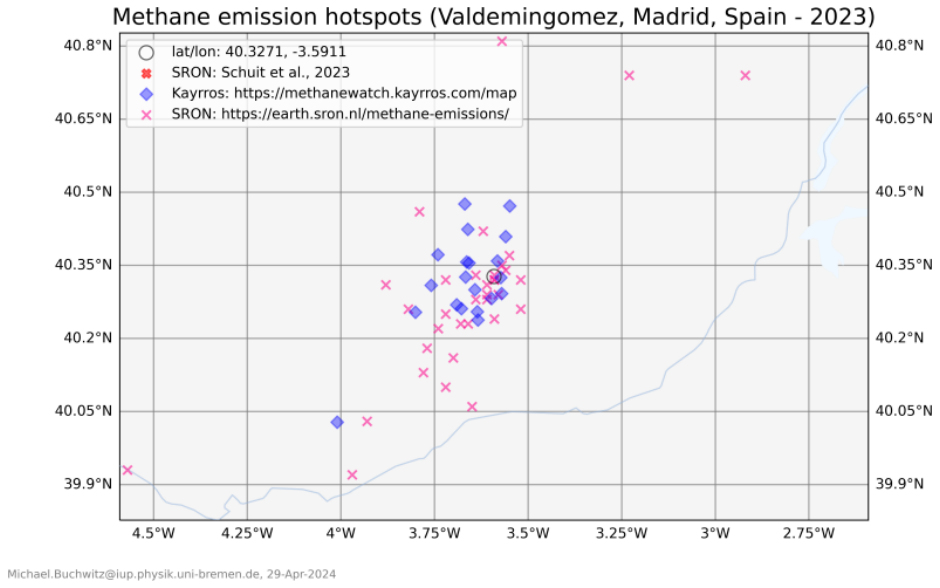


***Figure 4.1-1**: Locations of detected methane plumes near the Valdemingómez landfill in Madrid from TROPOMI/S5P via algorithms developed by SRON (Schuit et al., 2023, and https://earth.sron.nl/methane-emissions/) and Kayrros (https://methanewatch.kayrros.com/map) during (a) 2021 and (b) 2023.*

4.2. Intercomparison results for target locations

In this section the intercomparison results for the 10 pre-defined target locations are presented.

The target region intercomparison is based on 3 data sets, namely the data sets of estimated methane emissions as derived from TROPOMI/S5P using the algorithms from SRON, Kayrros and IUP-UB (see **Sect. 3.1**).

Figures 4.2-1 to 4.2-10 show detailed comparison results for each of the 10 target locations. The results from these comparisons are summarized in **Table 4.2-1**.

As can be seen, the IUP-UB algorithm provides typically the largest number of emission estimates and Kayrros the smallest (see Ndata in **Table 4.2-1**). The reason for this is the following: In contrast to the SRON and Kayrros algorithms, the IUP-UB CSF algorithm has been applied on all data with sufficient observational coverage even if there is no clear emission plume visible in the atmospheric methane map as plume identification is not part of this algorithm. If the IUP-UB algorithm is applied to a target location without emission sources (and therefore without emission plumes present in the analyzed maps) then the estimated emissions scatter (because of noise) around zero (with also negative emission estimates possible for single overpasses due to the noisy XCH_4 input data). The SRON and Kayrros algorithms are based on detection and quantification, and not all detections lead to quantification. As shown in this document, Kayrros uses a more conservative detection method, leading to less detections compared to SRON. Kayrros uses for quantification an inversion algorithm which is based on detailed modelling of emission plumes in contrast to the SRON IME algorithm. The requirements for successful detailed modelling are much higher compared to plume identification as plume modelling can be challenging due to limitations of the knowledge of the wind field, turbulence, etc. This is another reason why less emission estimates are available in the Kayrros data set compared to SRON.

In this context: As can be seen from **Table 4.2-1** and **Figures 4.2-9** and **4.2-10** only IUP-UB provides emission estimates for targets Karaturun and Hambach. As shown in **Figure 4.2-11**, which shows the average methane map for Hambach, there is no mean plume (in contrast to, for example, target Norte III, see **Figure 3.1.3-2**) but only a local apparent enhancement directly over the open pit coal mining area. Additional analysis, not shown here, indicates that retrieval conditions are very challenging directly over the coal mine due to complex topography combined with collocated variations of surface reflectance. This apparent methane increase is therefore most likely a retrieval artifact and not or only partially related to methane emissions. This finding (i.e., no clear emission plumes present) is consistent with the results from SRON and Kayrros as they do not provide emission estimates for this target.

For the elevated TROPOMI XCH_4 around Hambach in the German Rhineland area as reported by Ember DUH, 2024, see in particular this “Addendum: Satellite sees Germany’s CMM” which has been added to their website (https://ember-energy.org/latest-insights/de-undermines-cmm-emissions/) based on feedback from SRON: “The TROPOMI data indicated strong methane enhancements over the lignite

mines. Further analysis by the SRON Netherlands Institute for Space Research found that the enhancements are caused by high surface reflectivity and are not correlated with wind direction. Due to this limitation, we cannot conclude how much methane is present over the mines using the TROPOMI data. Observations from other satellite sensors and ground-based observations would help to quantify this.". The analysis presented here is consistent with this feedback from SRON to Ember DUH.

For target Karaturun, i.e., for Kazakhstan's Karaturun East oil field, massive methane emissions have been reported in Guanter et al., 2024, originating from a well blowout in 2023. For 2021, which has been investigated here, no emissions are reported by SRON and Kayrros in their corresponding data sets used here for comparison. As shown in **Figures 4.2-9** the IUP-UB analysis contains 7 overpasses in 2021 with good conditions (quality flag = 0 = OK) mostly scattering around zero (-17 +/- 28 t/h, i.e., the estimated emission is not significantly different from zero). These results indicate that in 2021 there were no major emissions exceeding the TROPOMI/S5P single overpass detection limit of about 8 t/h (e.g., Schuit et al., 2023).

As can also be seen, the number of collocations is typically small. Each algorithm has its own quality filtering (or data selection) procedure and shown here are only those estimates which are considered "good" for each of the 3 algorithms. **Table 4.2-1** shows that for the SRON – Kayrros comparison only for the Norte III target differences have been quantified as only for this target more than 2 collocations have been identified. Between IUP-UB and Kayrros no target location exists with more than 2 collocations.

In contrast, differences for several targets have been quantified for IUP-UB and SRON, namely for 7 of the 10 target locations. As can be seen, absolute emission differences are not constant but depend significantly on the target location. Mean differences and standard deviations of the IUP-UB and SRON differences are typically some 10 tons/hour with the largest difference for the Shanxi target, where differences exceed 100 tons/hour. The Shanxi area features steep elevation gradients and pooling of methane within valleys, complicating the application of mass balance emission quantification methods. Also, the relative differences vary significantly. Here relative differences are computed relative to the mean difference so that 67% corresponds to a ratio of 2. As can be seen, relative differences are on this order, indicating that a difference of a factor of 2 is quite typical, as can also be seen from **Figures 4.2-1 to 4.2-10**. But one has to note that essentially only a small number of points are collocated and used for comparison and that the standard deviation of the difference is often larger than the mean difference.

Despite the low number of collocations, **Table 4.2-1** shows in columns 2-4 ("Emissions 2021") that the average year 2021 emission estimates (as computed independently for each of the algorithms) often agree quite well, mostly within a factor of 2 and sometimes even significantly better (e.g., Hassi Messaoud, Korzephe, Norte III, Hail Creek).

However, for some targets there are also significantly larger differences, e.g., for target Zory, Upper Silesia, Poland. Here the difference even exceeds a factor of 3

(see also **Figures 4.2-5**). For this target there were, however, only 3 emission estimates made by IUP-UB in the first half of the year and only 4 detections by SRON mostly in the second half of the year starting approximately 2 months after the last estimate from IUP-UB and it is unclear to what extent emissions are constant during 2021. This target is one of the 3 targets where no collocated data exist.

Figure 4.2-6 shows the comparisons results for the Hail Creek open pit coal mine in Queensland, Australia. Note that this coal mine was also the topic of a recent aircraft-based publication (Borchardt et al., 2025) in which also IUP-UB CSF-based emission estimation results from TROPOMI/S5P are shown based on the method of Schneising et al., 2020b. The satellite-based emission estimates as reported in Borchardt et al., 2025, are (in that paper the reported uncertainties are 2-sigma and here divided by a factor of 2 for comparison with the 1-sigma values reported in this document): 25.0 ± 6.6 t/h (for 2018-2023) (Schneising et al., 2020b); 26.3 ± 2.9 t/h (for 2018-2019) (Sadavarte et al., 2021); 4.9 ± 1.2 t/h (for 2018-2019) (Palmer et al., 2021). While the IUP-UB (i.e, Schneising et al., 2020b) and SRON (i.e., Sadavarte et al., 2021) values agree very well, the Palmer et al., 2021, estimates are much lower. The reason for this is unclear. As shown in **Figure 4.2-6**, for 2021, the IUP-UB estimate for 2021 as obtained with the CSF v1.6 method used here for MEDUSA, is 27.5 ± 32.3 t/h (N = 24), which agrees well with the IUP-UB and SRON values reported in Borchardt et al., 2025 (but note the large standard deviation). The corresponding estimates (see **Figure 4.2-6** bottom) of SRON are 23.7 ± 6.6 t/h (N = 8) and the Kayrros estimate (N = 1) is 18.2 t/h. For Hail Creek also a figure similar as **Figure 3.1.3-2** has been generated showing that the average value of the emission as obtained using CSF v1.6 during 2018 – 2023 is 24.6 ± 4.3 t/h. This shows the IUP-UB and SRON emission estimates for Hail Creek are in good agreement even when different time periods are compared. However, the aircraft-based estimates as reported in Borchardt et al., 2025, are approximately a factor of 2 lower (depending on method and period they are in the range 9.6 – 14.0 t/h). These aircraft measurements have been obtained from relatively short campaigns, and it remains unclear to what extent they are representative for longer time periods.

Figure 4.2-7 shows the comparisons results for the area around Pecos located in the US Permian basin. The Permian basin is a major US source region for methane with many local methane sources mostly related to oil and gas exploitation and therefore many recent studies focused obtaining information on methane emissions in that area. For example, IUP-UB used the CSF algorithm described in Schneising et al., 2020b, to obtain emissions estimates using TROPOMI for the entire Permian basin and the results were in very good agreement with the independent study of Zhang et al., 2020, which had been carried out in parallel and also used TROPOMI data but a very different inversion method based on detailed modelling. Schneising et al., 2020b, summarize the comparison results as follows: “Concurrent with our study, Zhang et al. (2020) also quantified methane emissions from the Permian Basin using a different data set and an alternative inversion method combining information from the operational TROPOMI methane product and prior emission estimates within a Bayesian framework. Despite these quite distinct approaches, their total emission

estimate of 2.9±0.5 Mt/yr based on satellite observations from May 2018 to March 2019 agrees within uncertainties with our estimate. If we restrict our analysis to this specific period, the consistency becomes even better and we get the almost identical estimate of 2.8 Mt/yr with our method, which is independent of prior knowledge. Therefore, the corresponding absolute results are considered very robust.". This indicates that the IUP-UB CSF algorithm as described in Schneising et al., 2020b, produces reliable results. Nevertheless, most of the IUP-UB results reported here have been generated with a CSF algorithm which differs somewhat from the algorithm described in Schneising et al., 2020b (but note that we also present some comparison results for these 2 IUP-UB CSF algorithms). Probably even more important for the interpretation of **Figure 4.2-7** is that Pecos is not an area with a well isolated single emission source but an area containing many other sources nearby. Whereas the SRON algorithm identifies emission plumes the IUP-UB algorithm is also applied even if there are several plumes in the analyzed images. This could be the main reason why the IUP-UB emission estimates are significantly higher compared to the SRON estimates. In any case, comparisons may be difficult to be interpreted for complex regions such as parts of the Permian basin with its many emission sources located nearby.

Figure 4.2-12 shows comparisons limited to collocated emission estimates. Shown are comparisons for the corresponding 7 target locations where collocations exist. Shown are the estimated emissions (top panel) and corresponding ratios of the estimated emissions (bottom panel). Also listed are a number of numerical values related to the comparison for each target region: the median of the ratios, the number of colocations N and (if N > 1) the standard deviation of the ratios.

As can be seen, there are only very few collocations between IUP-UB and Kayrros (blue dots and blue text). The ratios cover a large range from close to zero (0.13) to 3.44 (for region HAS).

The number of collocations between SRON and Kayrros (orange dots and orange text) is somewhat larger but still small. Several ratios are close to 1 indicating good agreement. The largest number of collocations (N = 5) is present for target region NOR (Norte III landfill). Here the median of the ratios 1.53 and the standard deviation of the ratios is 0.6. The individual ratios cover the range from 0.6 to 2.4.

The largest number of collocations is present for IUP-UB and SRON (green dots and green text). The ratios cover a large range of values from -2 (as one IUP-UB emission estimate is negative) to 6.1 (both of these extreme cases are for region VAL). For some of the regions the median is around 1 (HAS, VAL, NOR) indicating good agreement on average but for four regions the median is larger than 2 (KOR, HAI, PER, SHA) indicating that the IUP-UB estimates are significantly higher compared to the SRON estimates.

Figure 4.2-13 is also similar as **Figure 4.2-12** but shown in its bottom panel the percentage overlap of the 1-sigma error bars, where 0% means that the errors bars do not overlap and 100% means full overlap. As can be seen, also this quantity varies a lot, covering the entire range from 0-100%.

Table 4.2-2 summarizes the comparison results shown in **Figures 4.2-12** to **4.2-13**. As can be seen, the IUP-UB estimates are overall a factor of 2 larger compared to Kayrros or SRON (but as shown in **Figure 4.2-12** the median of the ratio can also be close to 1 depending on the target region). There are only 10 collocations for SRON and Kayrros but for these few collocations agreement is close to 1 (within some 10% difference). **Section 5** features a comparison of all collocated SRON and Kayrros quantifications, showing good agreement as well.

Figure 4.2-14 shows correlations and corresponding ratios of emission for IUP-UB and Kayrros relative to SRON for all target locations combined. Conclusions are the same as already drawn from the discussion above, namely that there is scatter for individual quantifications (which is to be expected using different quantification approaches and wind data sets) and that on average the IUP-UB estimates are approximately a factor of 2 larger compared to the collocated SRON (N = 65) and Kayrros (N = 10) estimates, which agree quite well with a Kayrros to SRON ratio of emissions of 0.9 +/- 0.4 (N = 5) (the average SRON to Kayrros emission ratio is 1.3 +/- 0.5 as shown in **Table 4.2-2**).

To further investigate this potential IUP-UB overestimation, the IUP-UB CSF algorithm has also been applied to the operational TROPOMI/S5P Level 2 data product (V02 with recommended quality filtering, i.e., using qa_value > 0.5) in addition to using the scientific WFMD v1.8 XCH_4 product as input (using all “good” data, i.e., using quality_filter = 0). The results are shown in **Figure 4.2-15**, which has a structure similar to **Figure 4.2-14**. The comparison of these 2 figures shows that there are much fewer collocations between IUP-UB and SRON when the IUP-UB CSF algorithm is applied to the operational XCH_4 product. The main reason for this is that the operational XCH_4 product is typically sparser compared to the WFMD product (see **Figure 4.2-16** showing selected XCH_4 maps) and this results in less results being flagged “good”. Apart from that, the IUP-UB to SRON ratio of emissions is only slightly smaller (1.8 compared to 2.0) indicating that the approximately factor of 2 higher emission estimated by IUP-UB compared to SRON is likely not dominated or due to the use of different Level 2 input data but must be related to differences of the emission algorithms, i.e., to differences of the Level 2 to Level 4 algorithms.

To what extent the results shown here for the few selected target regions are representative also for other target regions needs to be investigated. This is partly done in WP 200 (validation) also for results obtained with IUP-UB’s CSF v1.6 algorithm. For the comparisons between SRON and Kayrros global data sets of emission estimates are available and detailed comparison results are shown in **Sect. 5**. Here only one comparison between these 2 data sets is shown in **Fig. 4.2-17**. As can be seen, **Fig. 4.2-17** confirms that the SRON and Kayrros estimates agree very well on average (the SRON/Kayrros ratio is 1.0 ± 0.8 and the Kayrros/SRON ratio is 1.4 ± 0.9) but the individual differences scatter significantly: approximately half of the emission estimates differ by more than 50% (fraction of ratios >1.5 or <2/3: 53.5%) and approximately a quarter of the cases differ by more that a factor of 2 (i.e., by more than 100%) (fraction of ratios >2 or <0.5: 27.1%).

Unfortunately, the true emissions during the satellite overpass are not known for any of the 10 target locations. It is therefore not possible to conclude who is closer to the truth.

So far it was only possible to compare emission estimates with known emissions using controlled release experiments. Unfortunately, the releases emissions are below the approximately 8 tons/hour detection limit of TROPOMI. However, emissions estimated from Hyperspectral Imager (HI) such as EnMAP and PRISMA can be and have been compared with known emissions from controlled release experiments. Detailed HI emission comparisons are reported in ANNEX B to AIR but one results is shown here. **Figure 4.2-18** shows results from single-blind Phase 1 of the controlled methane release experiment as conducted in Casa Grande, Arizona, USA, by teams from Stanford University and University of Michigan during January – March 2025. Shown are comparisons of satellite-derived estimated methane emissions as obtained using the IUP-UB CSF v1.6 algorithm to PRISMA and EnMAP XCH_4 enhancements compared to released ("true") emissions as made available by the Stanford / Michigan teams to all satellite retrieval teams after they have submitted their estimates. As can be seen from **Figure 4.2-18**, the IUP-UB CSF estimates agree very well with the "true emissions" (especially the "green" estimates corresponding to "good" quality). This indicates that the IUP-UB CSF algorithm does not have a fundamental problem that may cause a significant high bias. For the results shown in **Figure 4.2-18** also v1.6 of the IUP-UB CSF algorithm has been used, i.e., the same CSF v1.6 algorithm as used for the other IUP-UB results shown in this document. However, there are several differences depending if the algorithm is applied to S5P or to HI: (i) The Level 2 input data are different (the HI results are obtained using a Matched Filter algorithm (e.g., Roger et al., 2024) resulting in methane enhancement maps). (ii) The CSF box size (see **Fig. 3.1.3-1**) is 100 km for TROPOMI compared to 0.5 km for HI. (iii) For TROPOMI the PBL average is used for wind speed in contrast to the HI, where the 10 m wind speed is used (as the methane emitted near the ground is assumed to be mostly located close to the surface near the source). So there are several algorithm differences mostly related to the different spatial scales and to what extent HI CSF results are relevant also for TROPOMI is unclear but wind speed is one potentially relevant candidate.

To further investigate the impact of wind speed we show in **Figures 4.2-19** to **4.2-22** comparisons between IUP-UB and SRON using different wind speeds for the IUP-UB emission estimates.

As the IUP-UB emission estimates are directly proportional to wind speed no new inversions have been performed but simply a scaling with the alternatively used wind speed. **Figures 4.2-19** is similar as **Figure 4.2-12** but limited to the IUP-UB versus SRON comparison. The mean value of the ratios is 2.0 ± 1.4 (R = 0.76, N = 65).

Figures 4.2-20 is similar as **Figure 4.2-19** but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by GEOS-FP PBL wind speed. As explained in **Sect. 3.1.1**, the SRON algorithm uses different wind data sets and corresponding information is contained in the SRON file containing

their methane emission estimates. This SRON file contains (i) ERA5 10m wind speed (Hersbach et al., 2020), (ii) GEOS-FP 10m wind speed (Molod et al., 2012), and (iii) GEOS-FP (Molod et al., 2012) PBL wind speed. The corresponding wind speed values (as compiled by SRON) are used here for scaling the IUP-UB methane emission estimates, which are directly proportional to the assumed wind speed, to investigate the impact of replacing the IUP-UB default wind speed (ERA5 PBL average) by alternative wind speeds.

As shown in **Figure 4.2-20**, the mean value of the ratio of the modified IUP-UB (i.e., using GEOS-FP PBL wind speed instead of ERA5 PBL wind speed) and the SRON emission estimate is 1.8 ± 1.3 (R = 0.79, N = 65), which is 10% lower compared to the IUP-UB default wind speed (i.e., ERA5 PBL). But this “improvement” in terms of better agreement is only true on average. There are several regions where the agreement gets significantly worse (e.g., target region VAL (1.56 compared to 1.02), NOR (1.55 compared to 1.15) and PER (3.80 compared to 3.09)). This is probably because ERA5 and GEOS agree well “on average” but less good at smaller scales.

We also computed the change when using 10-meter winds instead of the PBL average: **Figures 4.2-21** is similar as **Figure 4.2-19** but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by GEOS 10-meter (10 m) wind speed. The mean value of the ratios is 1.4 ± 1.0 (R = 0.83, N = 65), which is significantly lower compared to the IUP-UB default wind speed (ERA5 PBL) and reduces the difference from a factor of 2 (+100%) to only a factor of 1.4 (+40%). **Figures 4.2-22** is similar as **Figure 4.2-19** but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by ERA5 10 m wind speed. The mean value of the ratios is 1.6 ± 1.3 (R = 0.72, N = 65), which is also significantly lower compared to the IUP-UB default wind speed (ERA5 PBL) and reduces the difference from a factor of 2 (+100%) to a factor of 1.6 (+60%).

This shows that the difference to SRON could be significantly reduced by using the 10 m wind instead of the PBL average. As shown above, the average ratio of IUP-UB emissions to SRON emissions when replacing the PBL wind by the 10 m wind in the IUP-UB algorithm is 25% (2.0 / 1.6 = 1.25). This finding is consistent with Schuit et al., 2023, also effectively concluding that the difference in wind speed is approximately 25% (as can concluded from their wind speed scaling factors, which are 0.59 for the 10 m wind and 0.47 for the PBL wind (0.59 / 0.47 = 1.25); see also Ueff formula below when discussing **Figure 4.2-14**).

So, the expected maximum effect on the IUP-UB emission estimates is an approximately 25% reduction of the estimated emissions if the 10 m wind is used instead of the PBL average. However, using the 10 m wind cannot be justified as it is unrealistic to assume that methane emitted near the ground will typically stay very close to the Earth surface for distances up to of several 10 km away from the source. On the other hand, assuming perfect mixing in the PLB seems also not realistic in all situations. Probably a better approach would be to use a wind speed somewhat lower than the PBL wind speed, e.g., 10% lower, but not a factor of 2 lower, except perhaps for certain situations.

Which wind is the most appropriate is an important question. This aspect could be investigated using simulations. Initial simulations have already been carried out for CH_4 and CO_2. For CO_2 simulated CO2M satellite XCO_2 retrievals have been carried out for a scene near Berlin containing CO_2 emission plumes of several coal-fired power plants. The atmospheric and surface parameters (CO_2, meteorology, surface properties, …) have been provided by EUMETSAT (responsible for CO2M Level 1 to Level 2 processing) and where mostly based on CAMS with corresponding underlying true (model) power plant CO_2 emissions provided by Gerrit Kuhlmann (Empa, Switzerland) in the context of the CO2M Mission Advisory Group. The IUP-UB CSF algorithm has been applied to the XCO_2 fields to estimate the power plants CO_2 emissions using the ERA5 PBL average and the ERA5 10-meter wind. The comparison with the true emissions for power plant Jänschwalde showed that the estimated emission is 12% too high if the PBL wind is used and 20% too low if the 10-meter wind is used (the results for the other power plants were difficult to interpret quantitatively, e.g., because of overlapping plumes). These results indicate that the "optimal wind speed" may be located somewhere in between the 10-meter and the PBL average wind speed but the optimal value probably depends also very much on location and time due to differences in meteorology, emission source characteristics, etc. For methane work on using simulated data has also started (e.g., using simulations around the Vidra landfill in Bucharest, Romania, provided by Empa in the context of the ESA SMART-CH4 project, but also using simulations as generated by SRON). The analysis is ongoing, but all initial results indicate that there is no obvious problem with the IUP-UB CSF algorithm that leads, for example, to a significant overestimation of the estimated emissions.

From the results presented here for the selected target regions it seems that the IUP-UB TROPOMI/S5P methane emission estimates have often a high bias compared to SRON (and perhaps also w.r.t. Kayrros but here the number of collocations is very small). However, detailed comparisons for several other scenes as conducted by B. Dils (BIRA-IASB) in the context of MEDUSA WP 220 show on average very good agreement between the emissions estimates as obtained by IUP-UB, SRON and Kayrros: slope $\beta = 1.01$ ($N = 26$, $R^2 = 0.74$) for SRON vs IUP-UB, $\beta = 0.92$ ($N = 22$, $R^2 = 0.69$) for Kayrros vs IUP-UB and $\beta = 1.05$ ($N = 22$, $R^2 = 0.55$) for SRON vs Kayrros. This indicates that all three algorithms result in very similar emission estimates on average (as all slopes are close to 1) although (as indicated by the moderate R^2 values) the individual emission estimates vary significantly (by often several 10%).

Based on this we conclude that due to the large variability (scatter) of the differences of the individual emission estimates (the scatter is often larger than the mean or the median difference), the low number of collocations and the significant differences depending on which scene is analyzed, it is difficult to clearly identify if one algorithm has in general a significant systematic bias in terms of the estimated methane emission compared to another algorithm.

Finally, we show comparisons between the IUP-UB and the SRON methane emission estimates (similar as **Figure 4.2-14**) but with modified IUP-UB emission

estimates as obtained by replacing IUP-UB's default wind speed (ERA5 PBL average) by SRON's "calibrated" GEOS-FP PBL effective wind speed. This effective wind speed Ueff is computed as follows using this formula as used by SRON (see Schuit et al., 2023): $Ueff = 0.47*U_{PBL}+0.31$ (for Ueff in m/s, i.e., for 0.31 m/s converted to the unit used by IUP-UB, i.e., km/h). The results are shown in **Figure 4.2-23**. As can be seen, the agreement between IUP-UB and the SRON is much better now: The IUP-UB / SRON emission ratio is now 0.7 ± 0.5 (R = 0.79, N = 65) compared to 2.0 ± 1.4 (R = 0.76, N = 65) for the original IUP-UB data (see **Figure 4.2-14**). The main reason for this is that the used wind speed is reduced by approximately a factor of 2 and the IUP-UB emission estimate is directly proportional to the assumed wind speed.

Figure 4.2-24 is similar as **Figure 4.2-23** but here the ERA5 PBL wind has been "calibrated" using the same Ueff formula (i.e., $Ueff = 0.47*U_{PBL}+0.31$) instead of the GEOS PBL wind. As can be seen, the comparison results are similar as for **Figure 4.2-23**: The ratio is now 0.8 ± 0.6 (R = 0.76, N = 65). The slightly better correlation of 0.79 shown in **Figure 4.2-23** for the GEOS wind speeds compared to the ERA3 wind speeds is probably due to the fact that SRON uses also GEOS winds.

Despite the better agreement with SRON one cannot conclude from **Figures 4.2-23** and **4.2-24** compared to **Figure 4.2-14** that the IUP-UB emissions with SRON's Ueff formula applied are more accurate compared to the original IUP-UB estimates (shown in **Figure 4.2-14**). SRON's Ueff formula has been obtained for SRON's IME algorithm and to what extent it improves IUP-UB's CSF algorithm estimates is unclear. This means it remains unclear which algorithm produces the more accurate results. Note that SRON's IME Ueff is an effective wind speed in the sense that it has no direct physical meaning. Only the ratio of "effective plume length" (L) divided by Ueff can be properly interpreted, namely as the time period needed by the emission source to emit the methane mass quantified as IME (IME in tons divided by time in hours given emission Q in tons/hour). Because only the ratio matters, both L (computed as square root of the identified plume area) and Ueff can be "wrong" but as long as the ratio L/Ueff is "correct", i.e., gives the right time period corresponding to IME. It therefore does not make sense to compare the wind speed as used by IUP-UB for the CSF method with the wind speed used by SRON for the IME method. In this context it is also important to point out that according to Varon et al., 2019, the use of the PBL average wind speed is appropriate for TROPOMI CSF-based emission estimates (their formula is Ueff = (1.05 ± 0.17)U_{PBL}, i.e., they recommend a scaling factor even a bit larger than 1.0). Based on this we conclude that the use of the PBL average wind speed seems a good choice for the TROPOMI CSF-based emission estimates.

From the above analysis we conclude that the use of different TROPOMI Level 2 input data and the use of the PBL average of the wind speed by IUP-UB are likely not the main reasons that can explain the factor 2 differences between the IUP-UB and the SRON results shown in **Figure 4.2-14**.

To further investigate this issue, parts of the IUP-UB algorithm have been modified to study sensitivities and to find out to what extent the CSF v1.6 algorithm can be improved. This resulted in a new version 1.7 of the IUP-UB CSF algorithm. The main differences of v1.7 compared to v1.6 are following:

- Modified background correction: Version 1.7 uses a simple smoothing approach, whereas for v1.6 it had been aimed at simulating transport of methane from outside the target region into the target region.
- Waiver of the 3-hour plume length limit: For v1.6 the maximum distance (length) of the plume from the target has been set to 3 hours. For v1.7 this criterion is not used any more. Consequently, all cross-sections up to a maximum distance of 100 km are used for v1.7.

In addition, several other minor changes have been implemented such as using a circular area around the target for initial data extraction rather than a rectangular region as used for v1.6.

To illustrate the changes, Figures **4.2-25** - **4.2-27** show comparisons between CSF v1.6 and v1.7 for selected satellite target overpasses using target region SHA (Shanxi) as example. As shown in Figures **4.2-8** and **4.2-12**, the IUP-UB CSF v1.6 emission estimates at target location SHA are typically significantly larger than the estimates of SRON. **Figure 4.2-25** shows the comparison for 19-April-2021. The v1.6 estimate (see panel (a)) is 346±117 t/h and the v1.7 estimate (see panel (b)) is 136±65 t/h, which is in much better agreement with the estimate of SRON, which is 157±49 t/h. The top left panels show XCH_4 around the target location and the top panel in **Figure 4.2-25 (b)** shows all extracted data in the surrounding area are now located in a circular area around that location as mentioned above. The top right and bottom left maps show the methane anomalies, ΔXCH_4, i.e., XCH_4 after background correction. As can be seen, the maps for v1.6 and v1.7 differ significantly, and this impacts the resulting emission estimates. As shown in the bottom middle figures, the number of used cross-sections (shown as red crosses) extend to 100 km for v1.7 as the 3-hour length limit is not used anymore for v1.7.

The modification with the largest impact on the emission estimate is the improved background correction.

However, also the waiver of the 3-hour limit typically results in a reduction of the reported emission estimate as typically the computed fluxes through the cross-sections are lower for cross-sections further away.

But it is also important to know how the reported emissions need to be interpreted. For v1.6, with the 3-hour limit implemented, the reported emission referred to the average emission in the period 3 hours before the satellite overpass to the time of the satellite overpass. For v1.7 this averaging time is not constant anymore but depends on wind speed. For an emission source emitting at a constant rate both approaches lead (ideally) to identical results but this is not true for sources with time dependent emission.

Figure 4.2-26 shows another example, namely the corresponding results for region SHA for 3-December-2021. Here, the v1.6 estimate is 262±66 t/h and the v1.7 estimate is 34±16 t/h, which is also in much better agree with the estimate of SRON, which is 46±16 t/h.

There are however also several cases where the v1.6 estimate is in better agreement with SRON. An example is shown in **Figure 4.2-27** corresponding to 5-May-2021. Here, the v1.6 estimate is 98±39 t/h, the v1.7 estimate is 62±39 t/h, and the SRON estimate is 103±45 t/h.

Figure 4.2-28 shows a comparison of all collocations with SRON for v1.6 and v1.7. **Figure 4.2-28** is similar as **Figure 4.2-14 (left)** but with CSFv1.7 results added (in red) to be compared with the CSF v1.6 results (in green). As can be seen, the CSF v1.7 results agree on average better with the SRON results compared to the CSF v1.6 results. The correlation coefficient increases moderately from 0.76 to 0.81 but the ratio changes from a factor of 2 overestimation (2.0 ± 1.4 for v1.6) to a 30% underestimation (0.7 ± 0.4 for v1.7).

These results show that the CSF algorithm depends critically on how specific aspects are addressed, in particular the background correction. At this stage only a limited amount of data has been processed with v1.7 and additional data processing and comparisons are needed to find out to what extent further changes are needed.

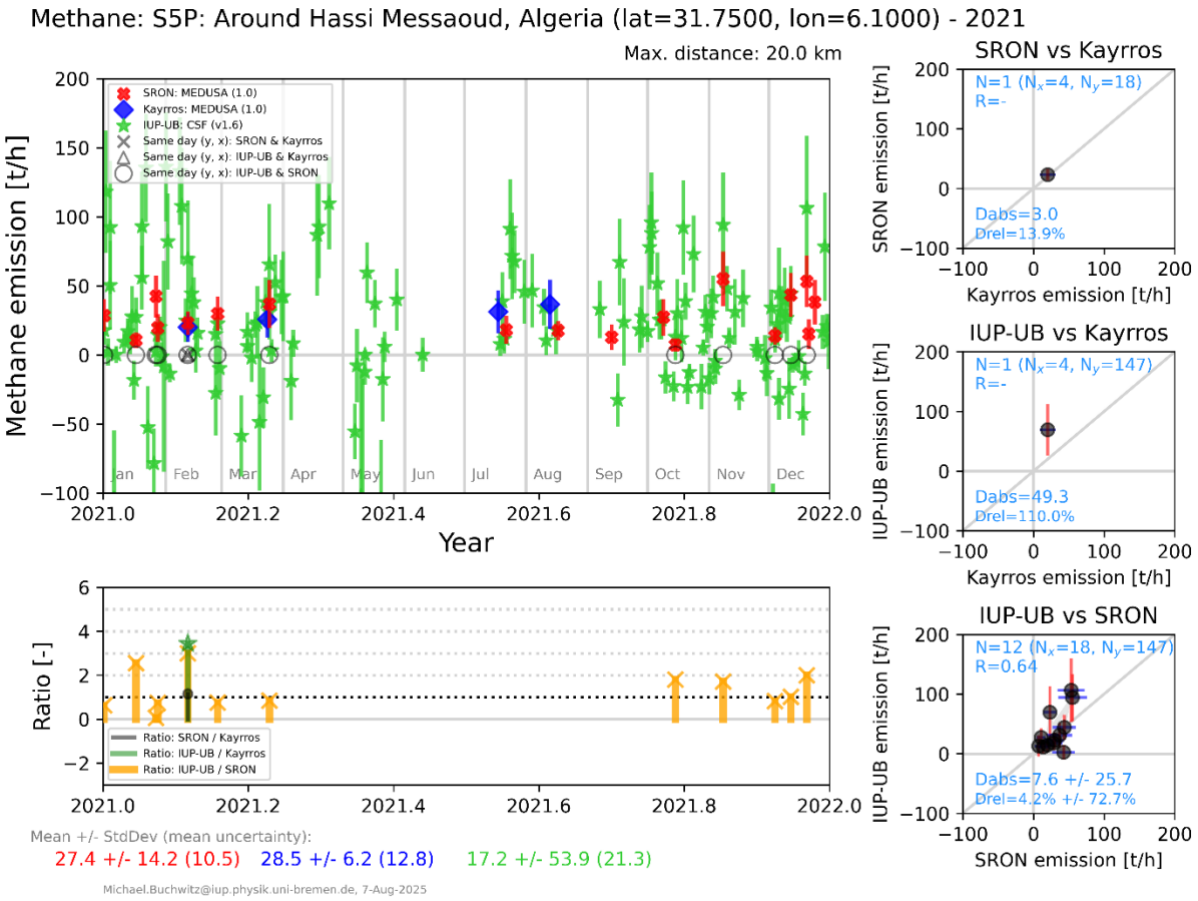


Figure 4.2-1: *Comparison results for target location HAS, i.e., Hassi Messaoud, Algeria. Top left: Estimated emissions including 1-sigma error bars. Bottom left: Ratios of emissions. Right: Correlation plots for algorithm pairs including error bars and numerical results (number of pairs (N) and individual estimates (Nx and Ny), linear correlation coefficient R, absolute difference Dabs (mean +/- standard deviation) and relative difference Drel (computed as relative difference w.r.t. the mean difference).*

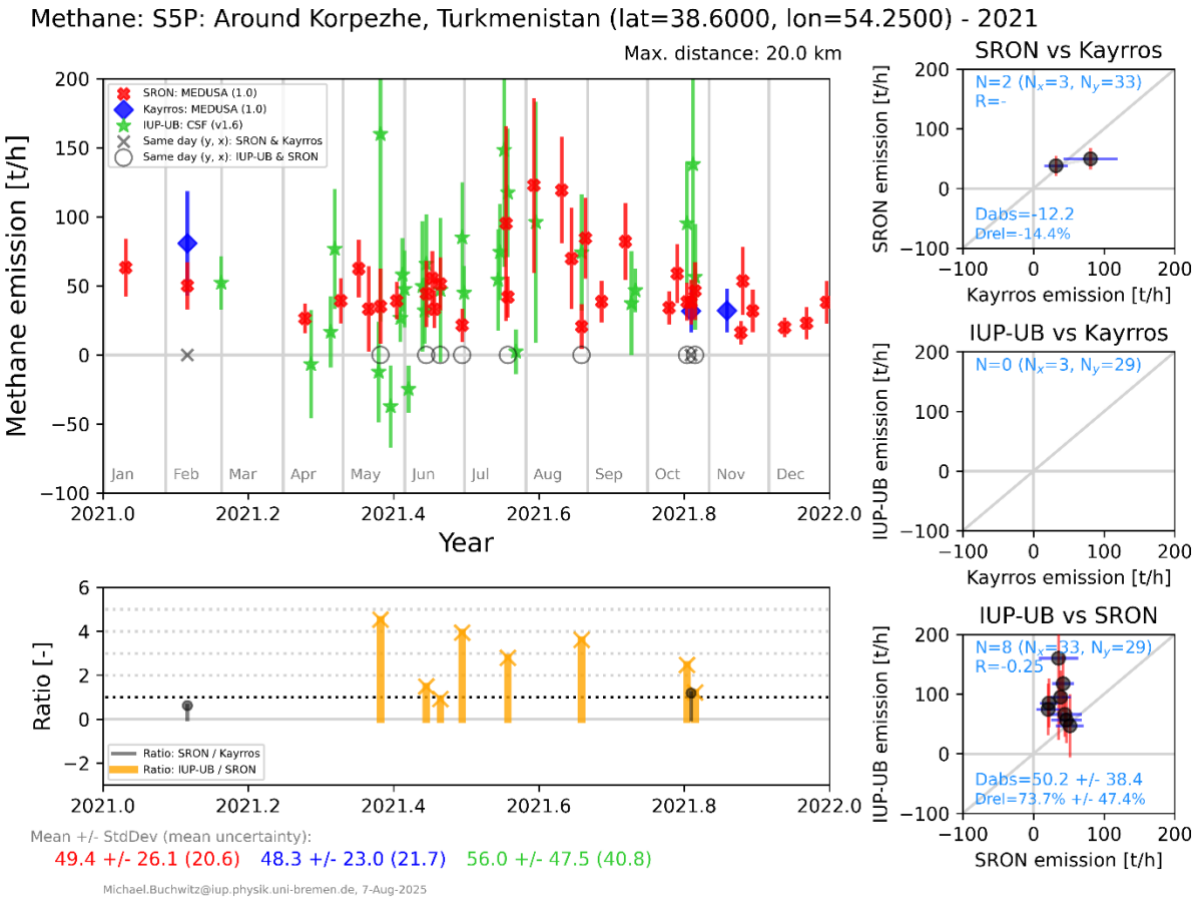


Figure 4.2-2: *Comparison results for target location KOR, i.e., Korpezhe, Turkmenistan.*

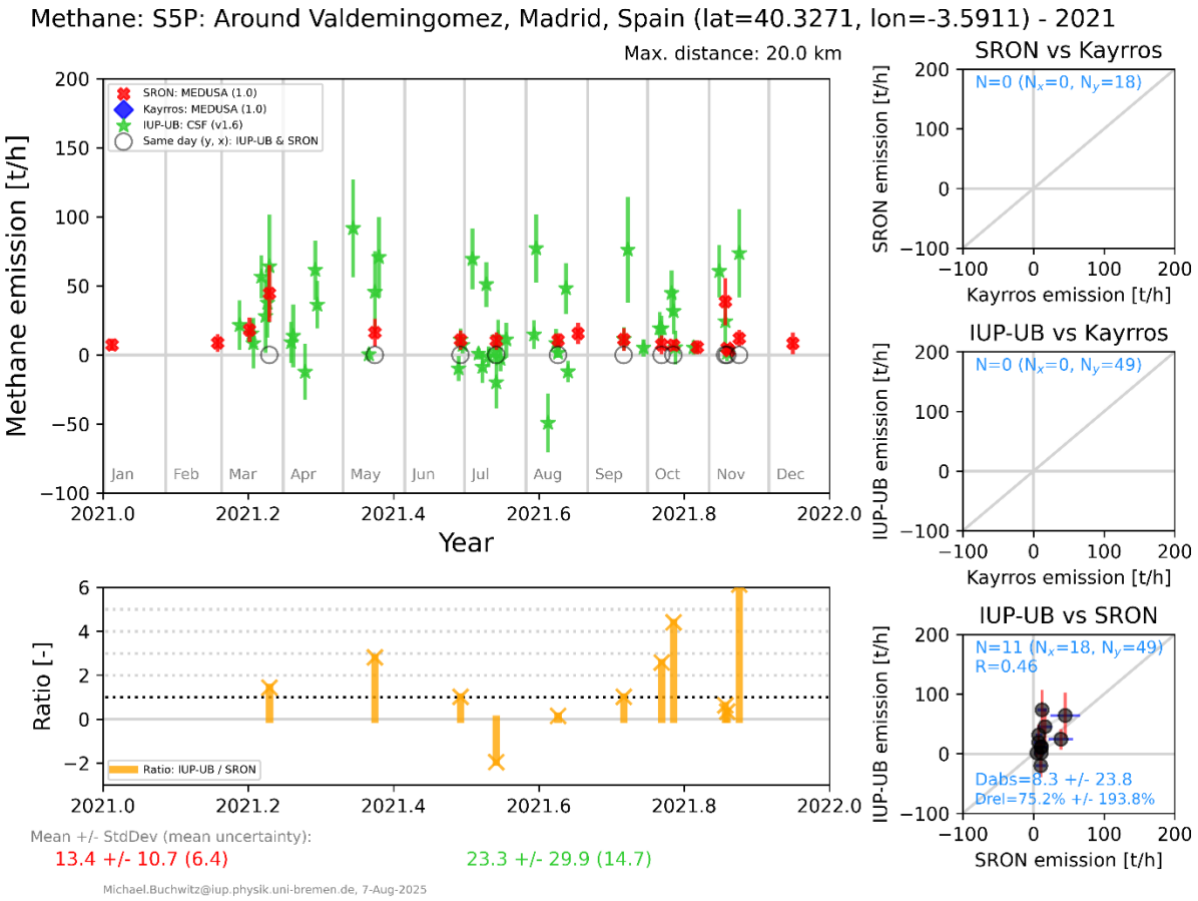


Figure 4.2-3: *Comparison results for target location VAL, i.e., for the Valdemingomez landfill, Madrid, Spain.*

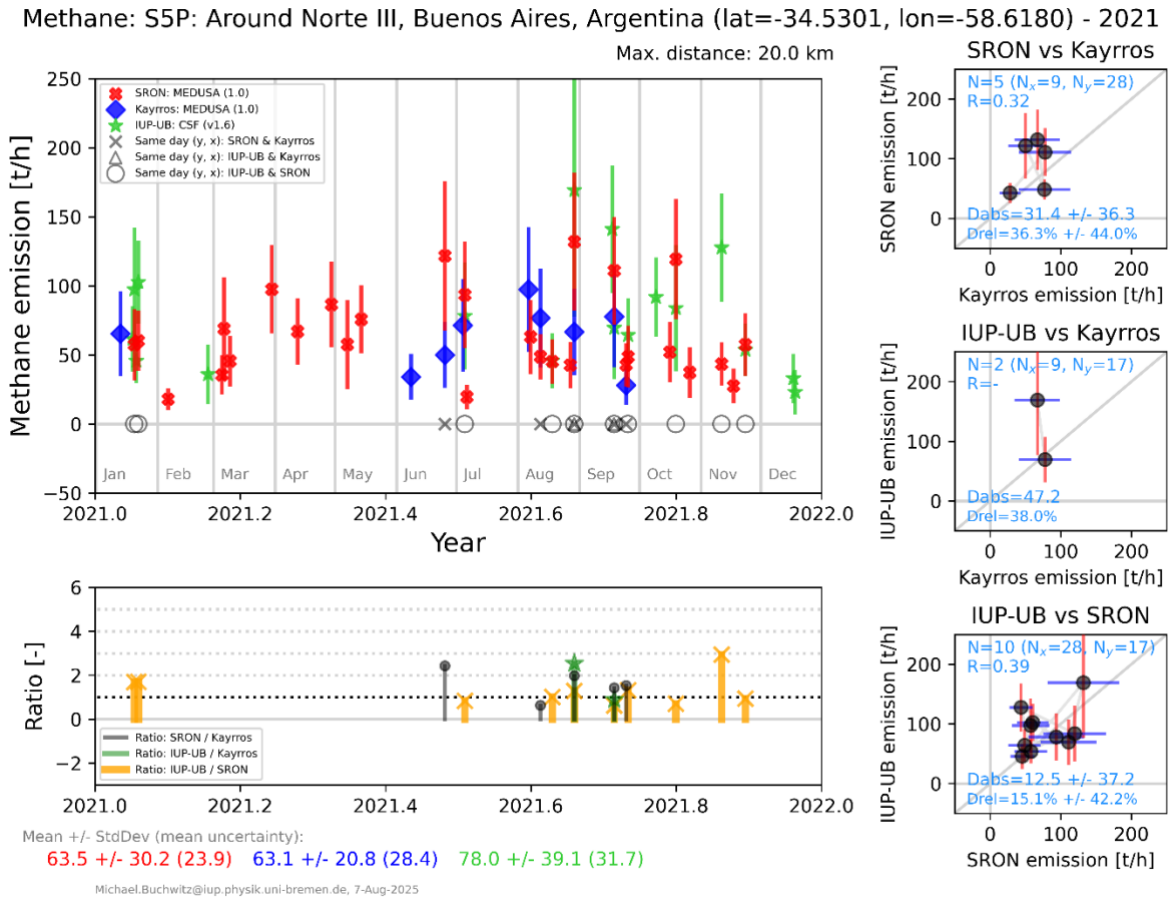


Figure 4.2-4: *Comparison results for target location NOR, i.e., for the Norte III landfill, Buenos Aires, Argentina.*

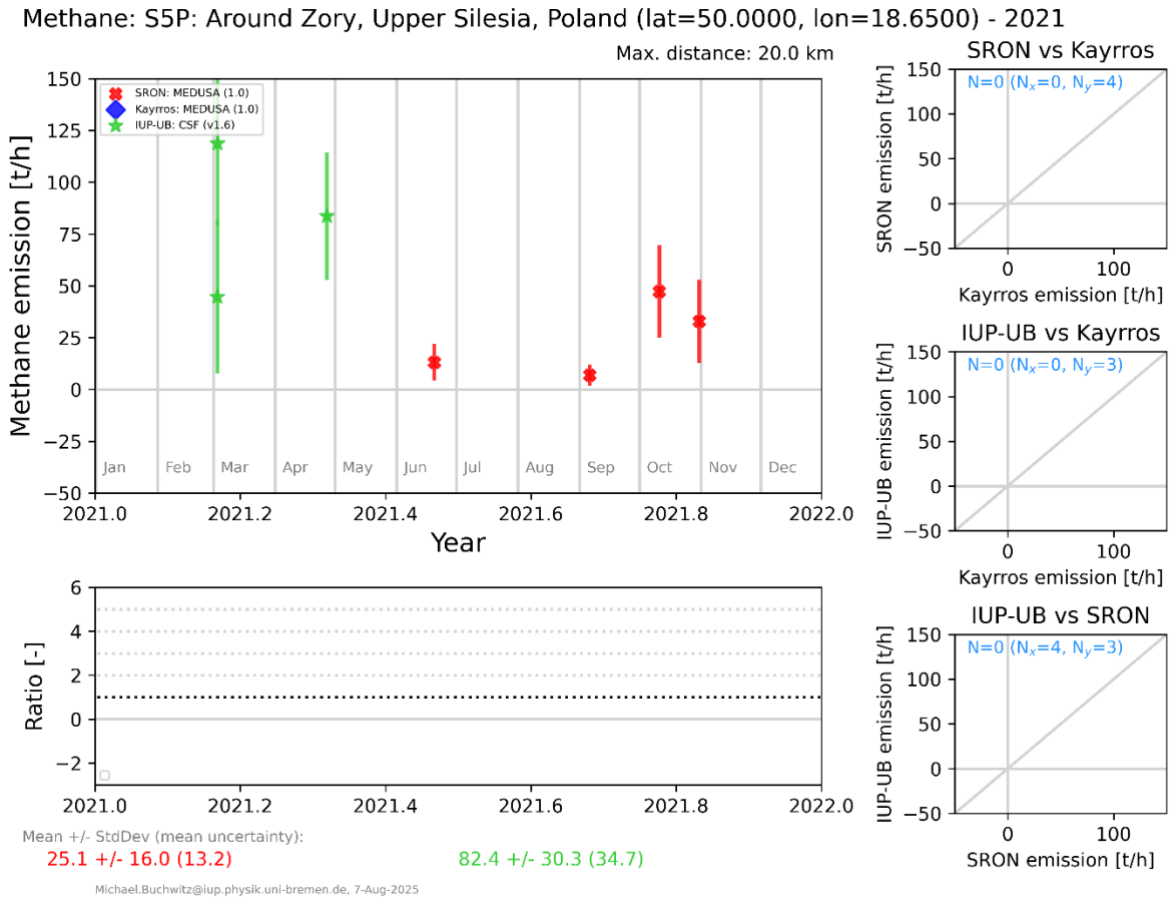


Figure 4.2-5: *Comparison results for target location USI, i.e., Zory, Upper Silesia, Poland.*

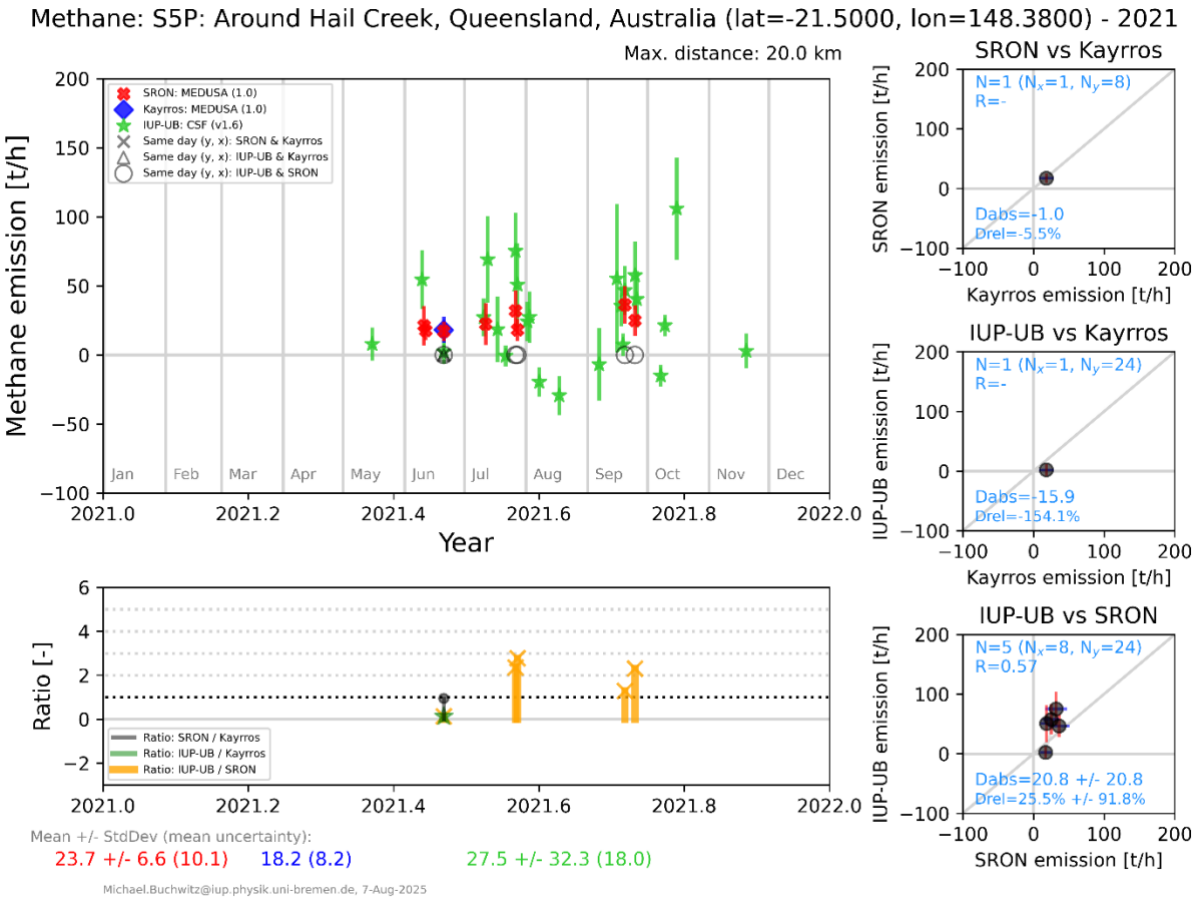


Figure 4.2-6: *Comparison results for target location HAI, i.e., Hail Creek coal mine, Queensland, Australia.*

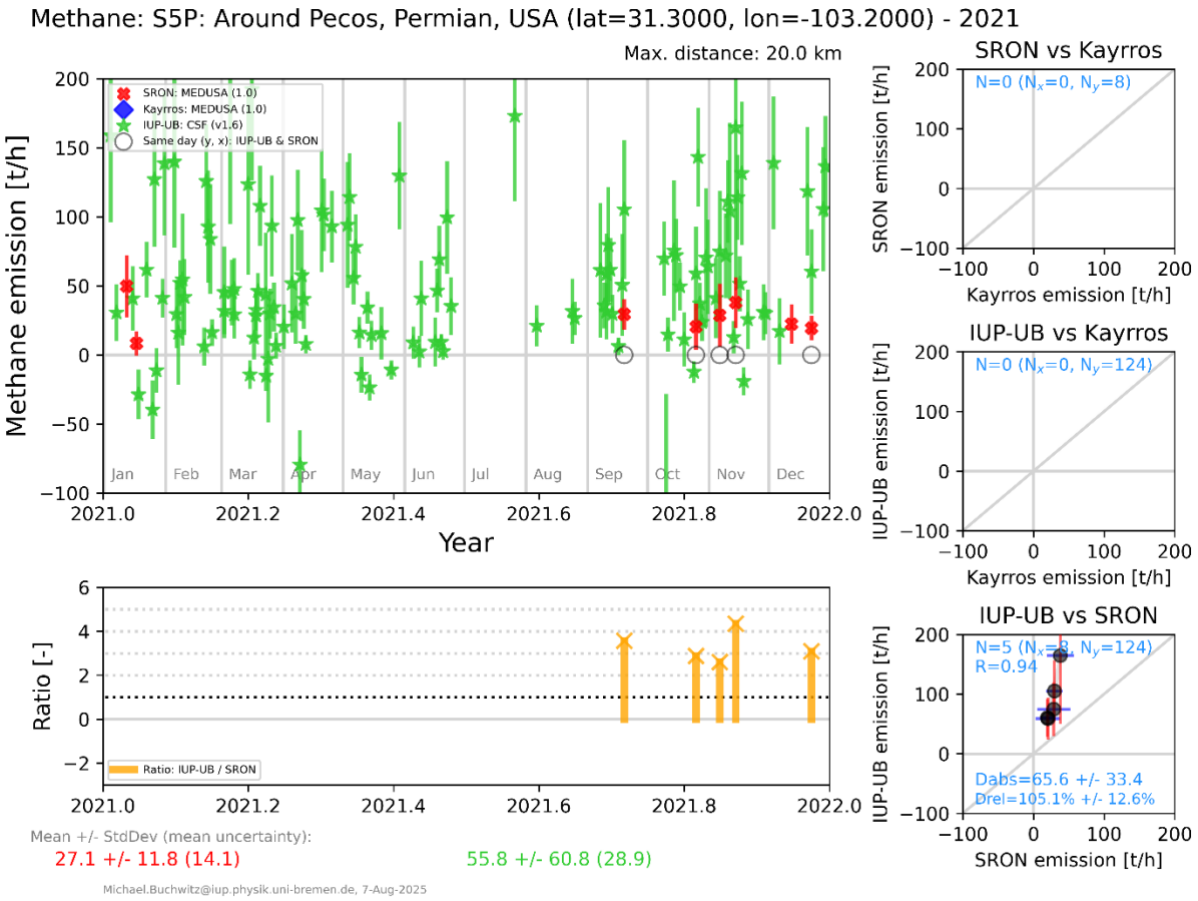


Figure 4.2-7: *Comparison results for target location PER, i.e., Pecos, Permian, New Mexico, USA.*

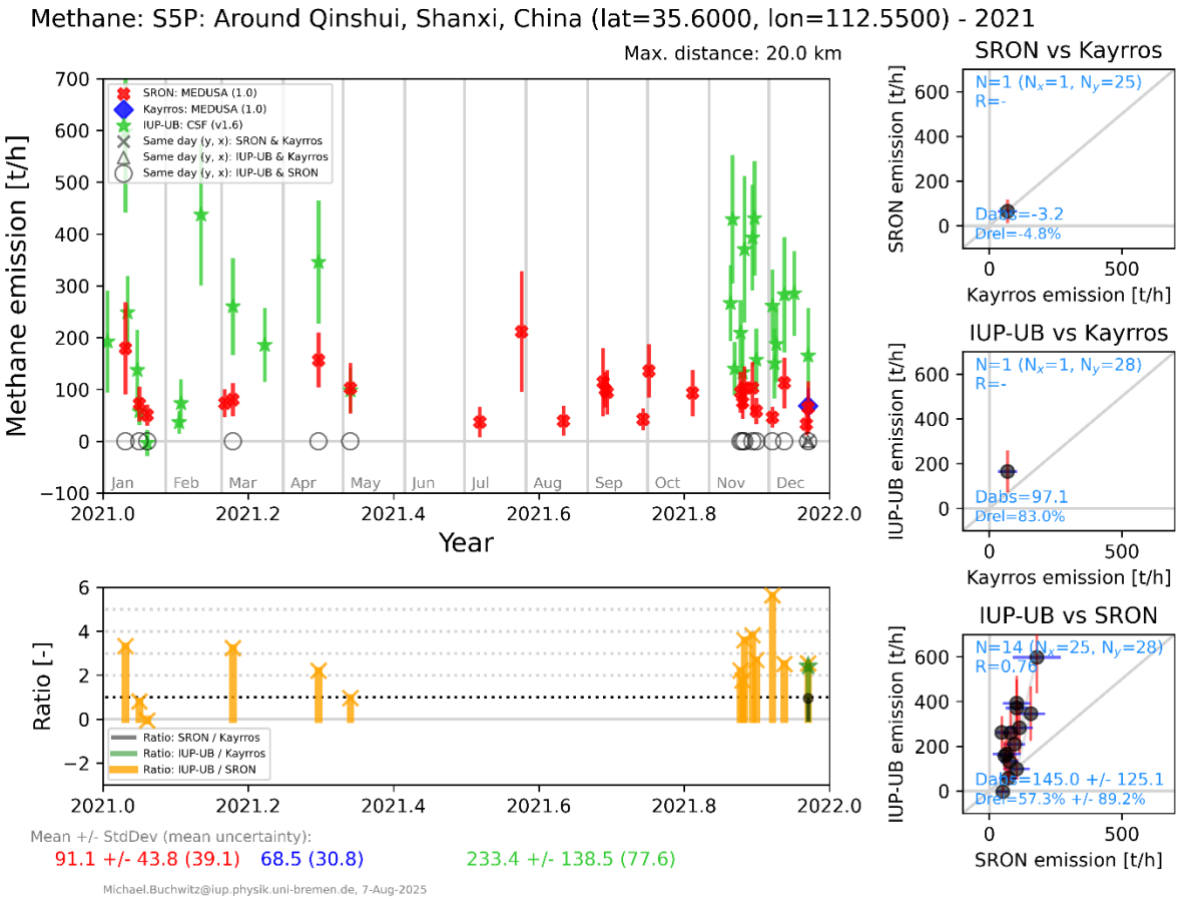


Figure 4.2-8: *Comparison results for target location SHA, i.e., Shanxi, China.*

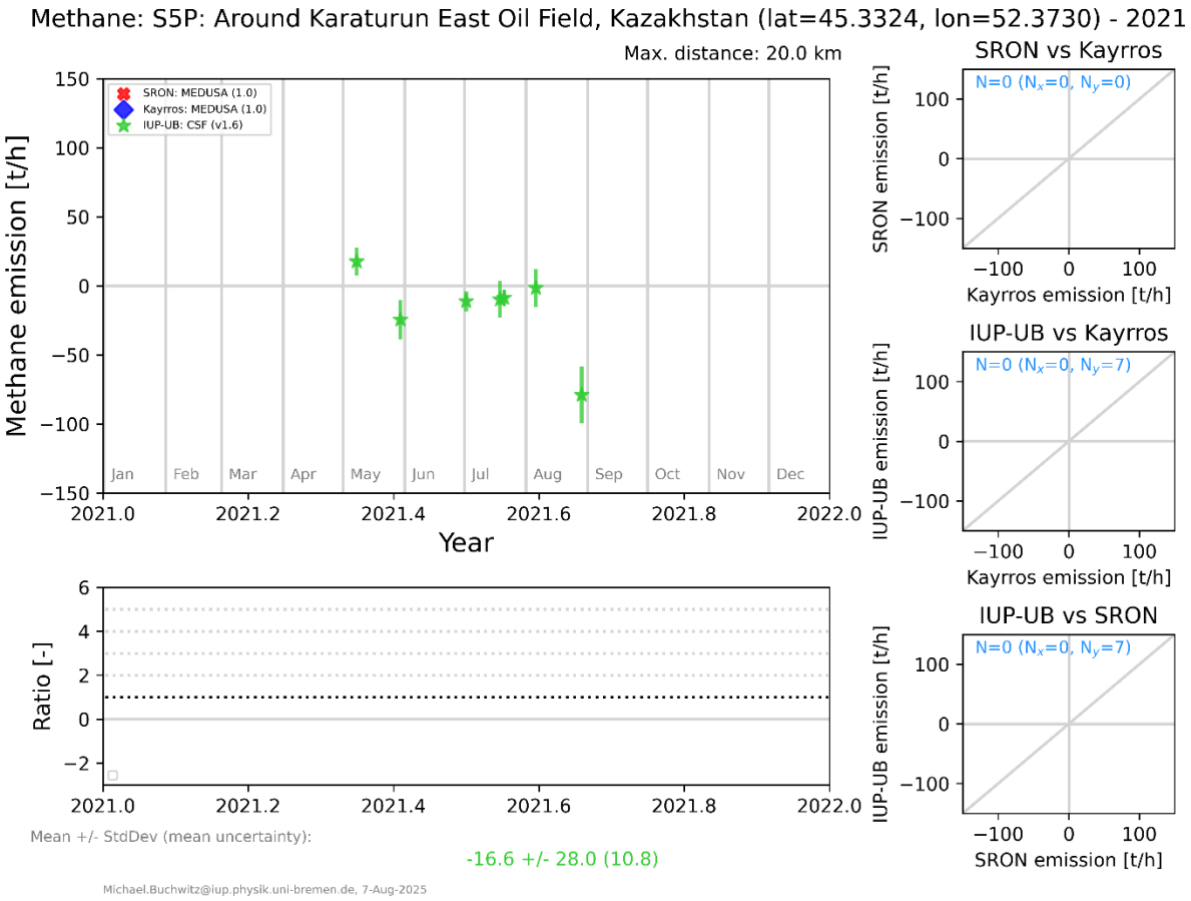


Figure 4.2-9: *Comparison results for target location KAR, i.e., Karaturun East oil field, Kazakhstan.*

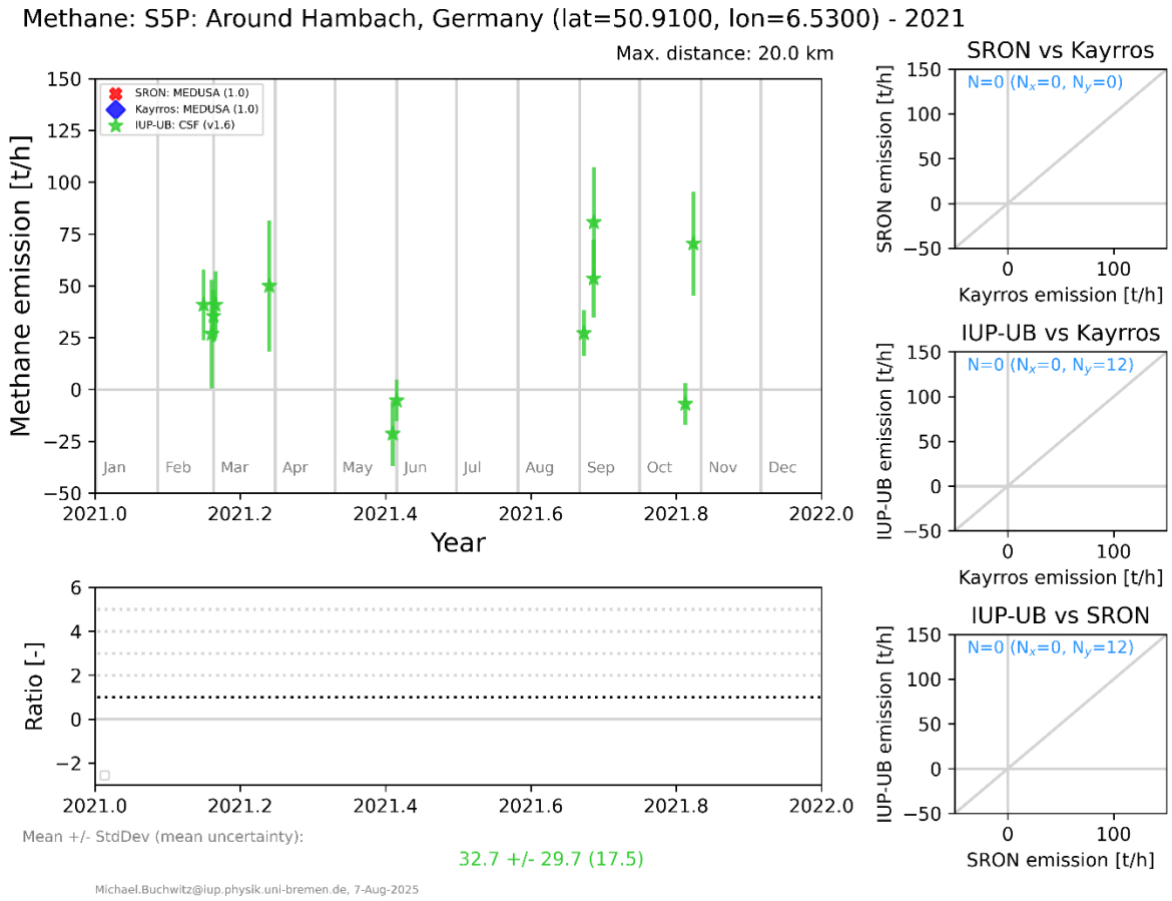


Figure 4.2-10: *Comparison results for target location HAM, i.e., near Hambach coal mine, Germany.*

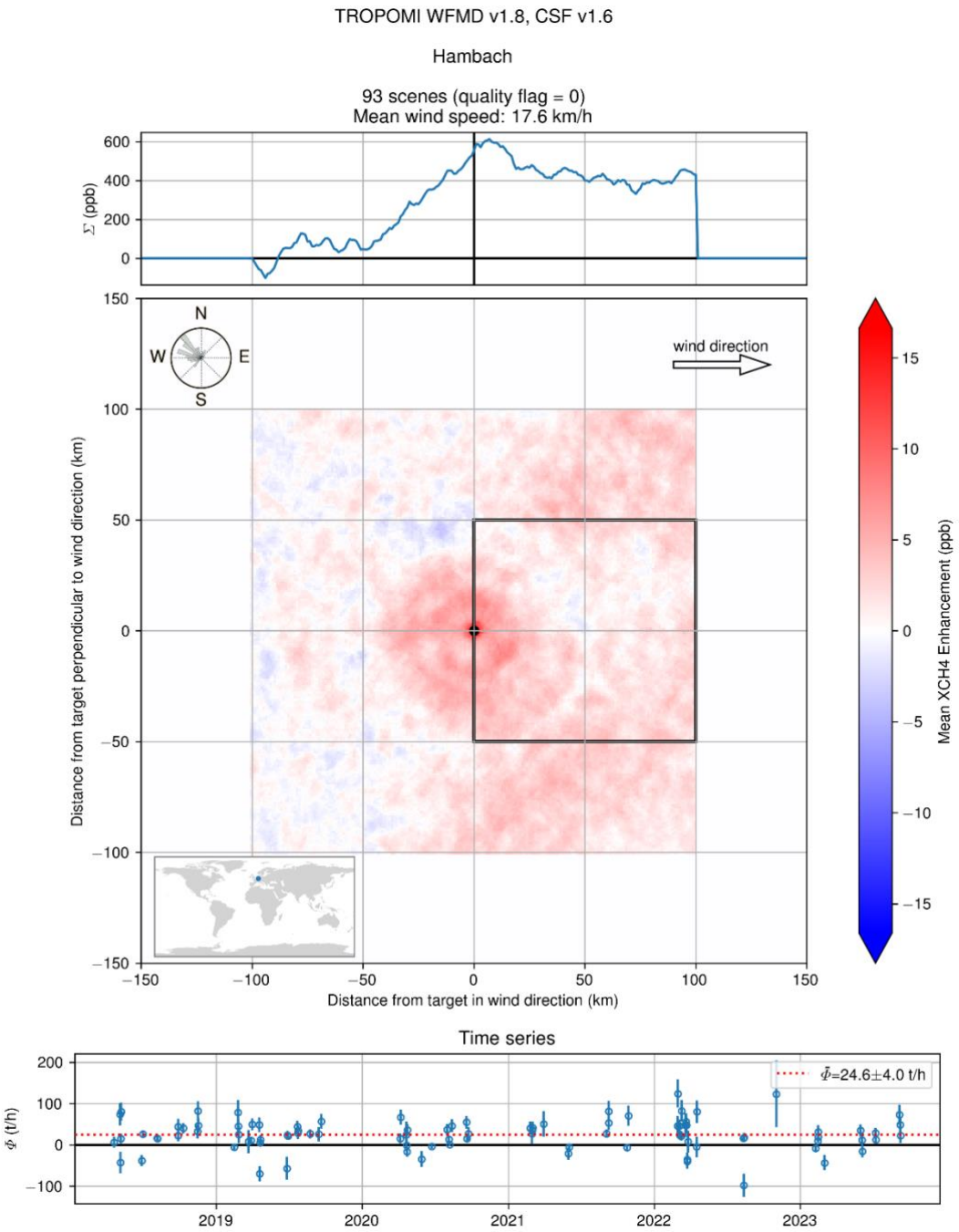


Figure 4.2-11: *As* ***Fig. 3.1.3-2*** *but for target location HAM, i.e., at and around the Hambach coal mining area in Germany.*

Target ID Short name	Emissions (2021) All in 2021; Mean ± Stddev [t/h] (Ndata)			Differences (#): Absolute [t/h] (Ncoloc) Relative Mean ± Stddev [%] (R)			Comments
	SRON	Kayrros	IUP-UB	SRO - Kay	IUP - Kay	IUP - SRO	
FM-01 Hassi_Messaoud	27 ± 14 (18)	29 ± 6 (4)	17 ± 54 (147)	- -	- -	8 ± 26 (12) 4 ± 73 (0.64)	
FM-02 Korpezhe	50 ± 26 (33)	48 ± 23 (3)	56 ± 48 (29)	- -	- -	50 ± 38 (8) 74 ± 48 (-0.25)	
FM-03 Valdemingomez	13 ± 11 (18)	- (0)	23 ± 30 (49)	- -	- -	8 ± 24 (11) 75 ± 194 (0.46)	
FM-04 Norte_III	64 ± 30 (28)	63 ± 21 (9)	78 ± 39 (17)	31 ± 36 (5) 36 ± 44 (0.32)	- -	13 ± 37 (10) 15 ± 42 (0.39)	
FM-05 Upper_Silesia_Zory	25 ± 16 (4)	- (0)	82 ± 30 (3)	- -	- -	- -	
FM-06 Hail_Creek	24 ± 7 (8)	18 (1)	28 ± 32 (24)	- -	- -	21 ± 21 (5) 26 ± 92 (0.57)	
FM-07 Permian_Pecos	27 ± 12 (8)	- (0)	56 ± 61 (124)	- -	- -	66 ± 33 (5) 105 ± 13 (0.94)	
FM-08 Shanxi_Qinshui	91 ± 44 (25)	69 (1)	233 ± 139 (28)	- -	- -	145 ± 125 (14) 57 ± 89 (0.76)	
FM-09 Karaturun_East	- (0)	- (0)	-17 ± 28 (7)	- -	- -	- -	No plumes in 2021
FM-10 Hambach	- (0)	- (0)	33 ± 30 (12)	- -	- -	- -	No plumes in 2021

Ndata: SRON & Kayrros: Number of plumes; IUP (WFMDv1.8, CSFv1.6): Number of overpasses with dense data coverage

(#) Only reported if more than 2 colocations (Ncoloc > 2) Relative difference [%] := (a-b)/mean (example: a=2, b=1 -> 67%)

Table 4.2-1: *Summary of the target region comparison. Columns in "Emissions (2021)" list the mean +/- standard deviation of estimated emissions computed independently for the three algorithms including number of data (Ndata). Emission differences are listed in the following 3 columns but only if the number of pairs of collocated data (Ncoloc) is larger than 2.*

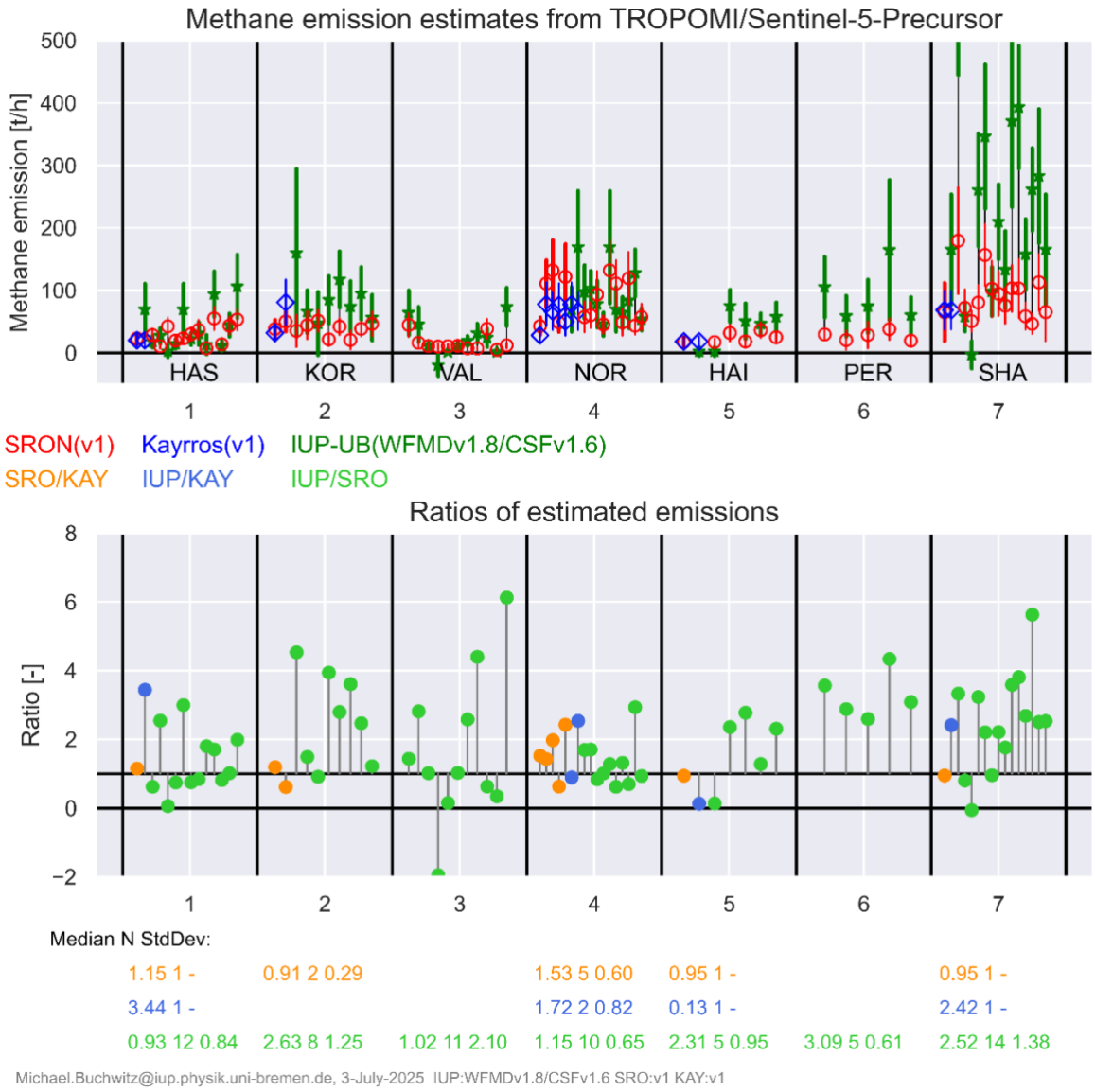


Figure 4.2-12: *Comparison of estimated emissions for target locations with collocated emission estimates (7 from 10). Top: Estimated emission. Bottom: Ratios of estimated emissions. For each of the 7 target regions the following numerical quantities are listed: the median of the ratios, the number of collocations N and (if N > 1) then standard deviation of the ratios.*

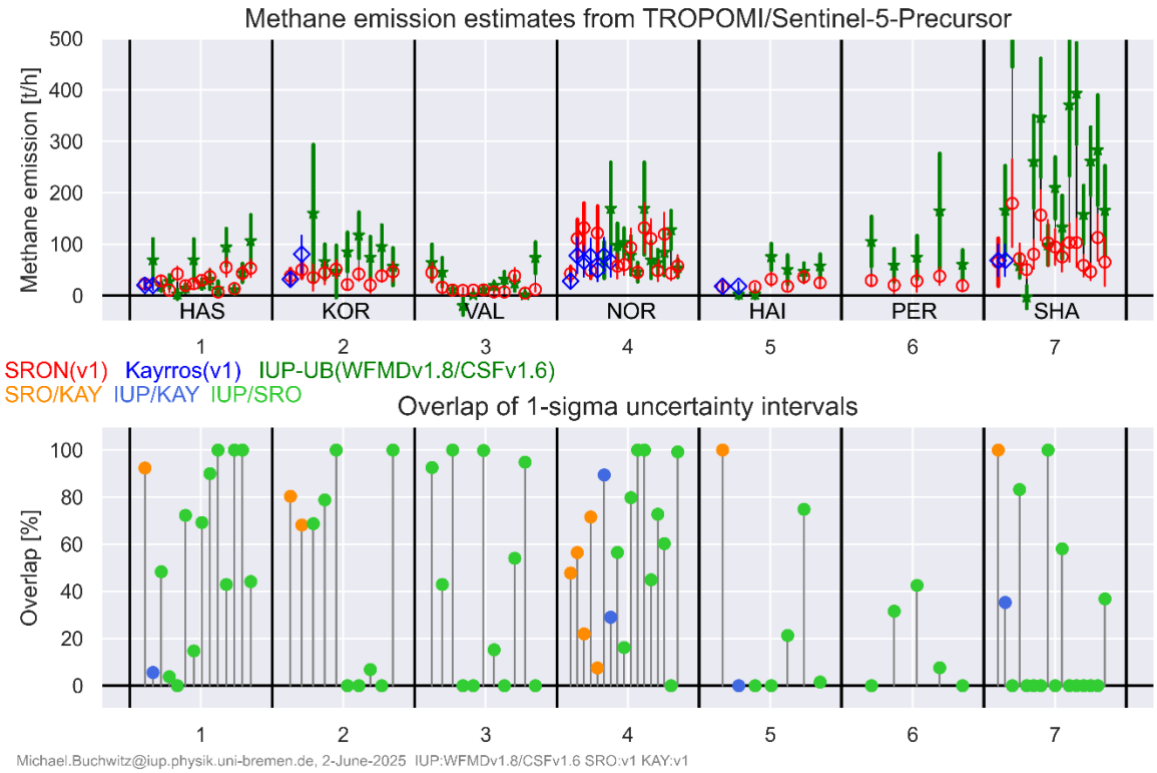


Figure 4.2-13: *Similar as* ***Figure 4.2-12*** *but with the overlap fraction of the uncertainty intervals in the bottom panel.*

Algorithms A & B:	SRON & Kayrros	IUP-UB & Kayrros	IUP-UB & SRON
N [-]	10	5	65
Ratio emissions A/B [-] Mean±StdDev (median)	1.3 ± 0.5 (1.2)	1.9 ± 1.2 (2.4)	2.0 ± 1.4 (1.8)
Ratio emissions B/A [-] Mean±StdDev (median)	0.9 ± 0.4 (0.9)	2.0 ± 2.9 (0.4)	0.8 ± 3.2 (0.5)
Emission difference A-B [t/h]	13 ± 33	45 ± 50	49 ± 84
Uncertainty overlap fraction [%]	65 ± 30	32 ± 32	40 ± 40
R [-]	0.56	0.68	0.76

Table 4.2-2: *Summary of the target locations comparison for collocated emission estimates. N is the number of collocations of the two algorithms listed in the first row for columns 2-4. "Ratio" are the ratios of the emission estimates in terms of mean and standard deviation and (in brackets) the median. "Uncertainty overlap fraction" quantifies the percentage overlap of the 1-sigma error bars (no overlap = 0%; one error bar entirely within range of other error bar = 100% overlap). "R" is the linear correlation coefficient.*

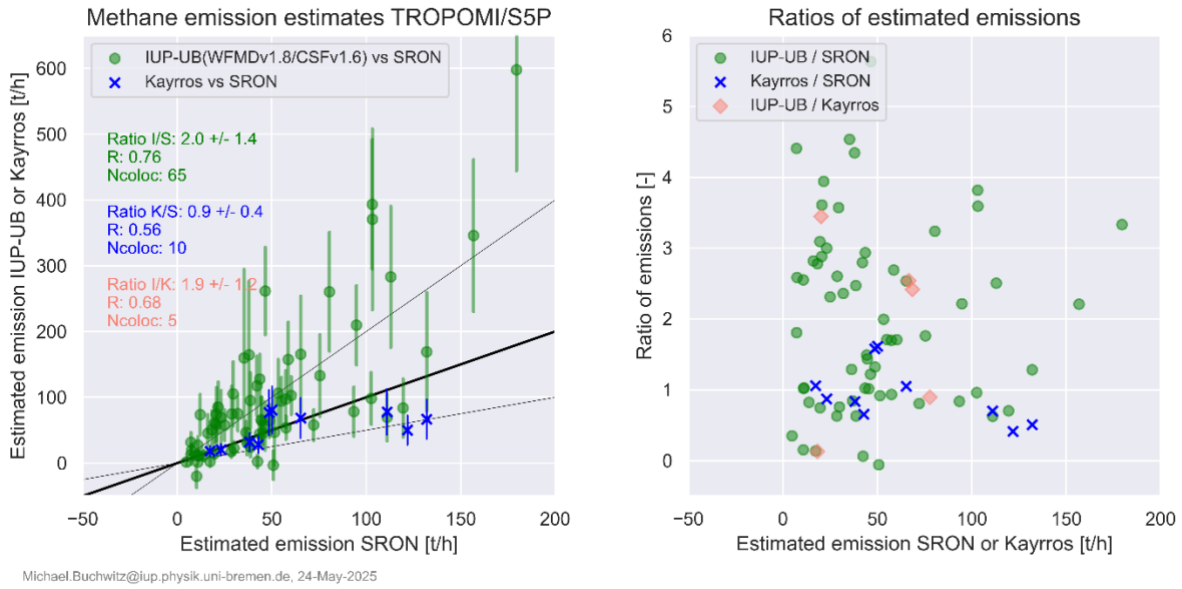


Figure 4.2-14: *Left: Comparison of collocated emission estimates (green: IUP-UB & SRON; blue: Kayrros & SRON; light red: IUP-UB and Kayrros (only the numerical results are shown for this pair of algorithms)). Listed are the numerical values of the following quantities: ratio of emissions in terms of mean value and standard deviation, linear correlation coefficient R, number of collocations Ncoloc. Right: corresponding ratios of emissions.*

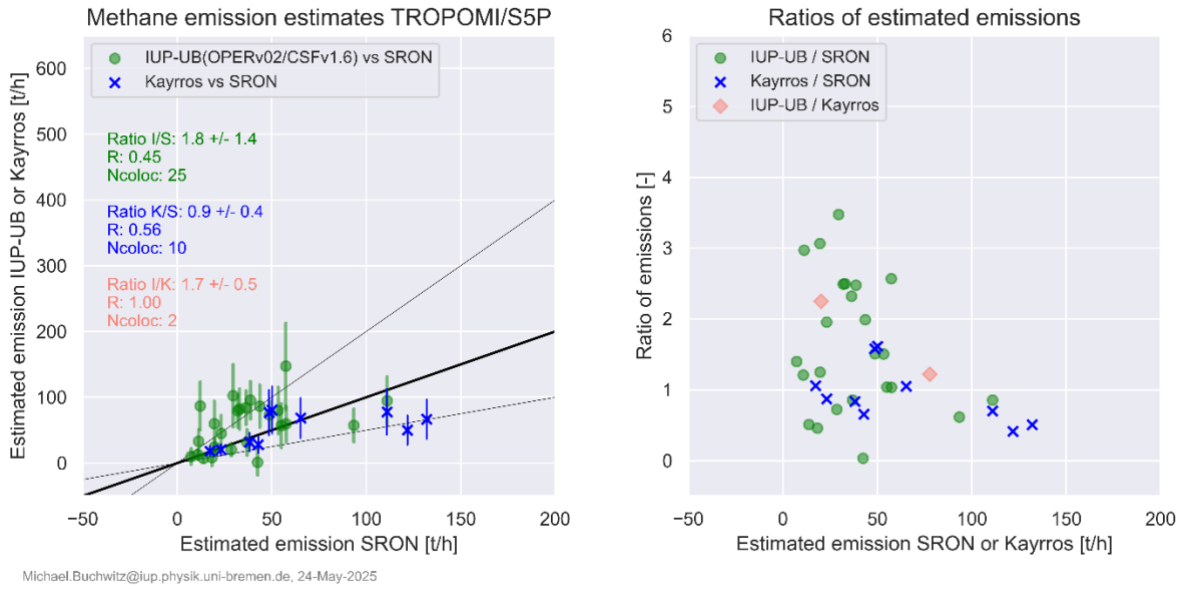


Figure 4.2-15: *As **Figure 4.2-14** but with the IUP-UB CSF algorithm applied to the operational TROPOMI/S5P XCH_4 product (V02) instead of the WFMD v1.8 Level 2 product.*

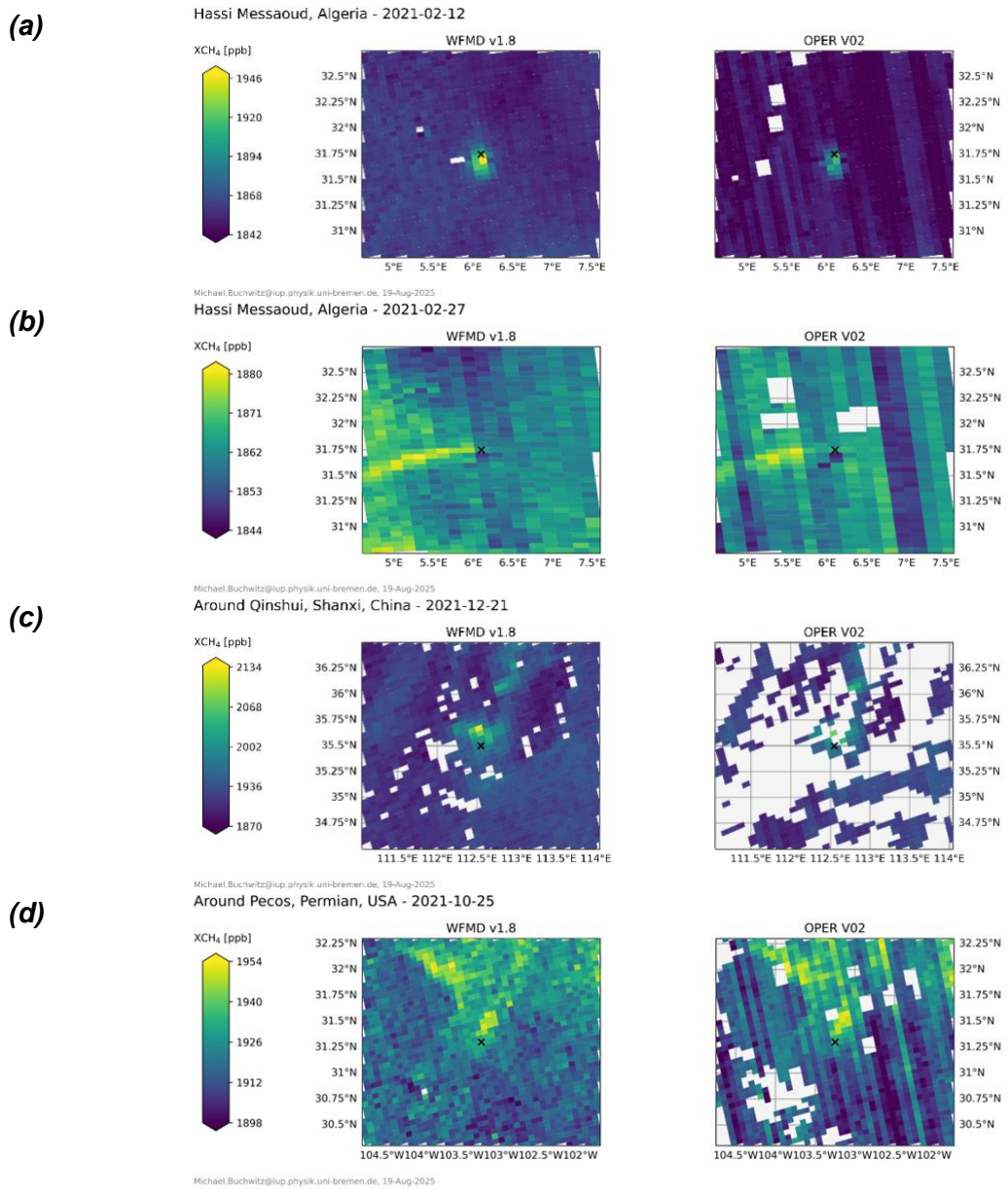


Figure 4.2-16: *Comparisons of satellite single overpass TROPOMI XCH4 maps as obtained with WFMD v1.8 (left) and the OPERational V02 (right) algorithms. (a) Target Hassi Messaoud on 12-Feb-2021, (b) Hassi Messaoud on 27-Feb-2021, (c) Shanxi on 21-Dec-2021 and (d) Permian on 25-Oct-2021.*

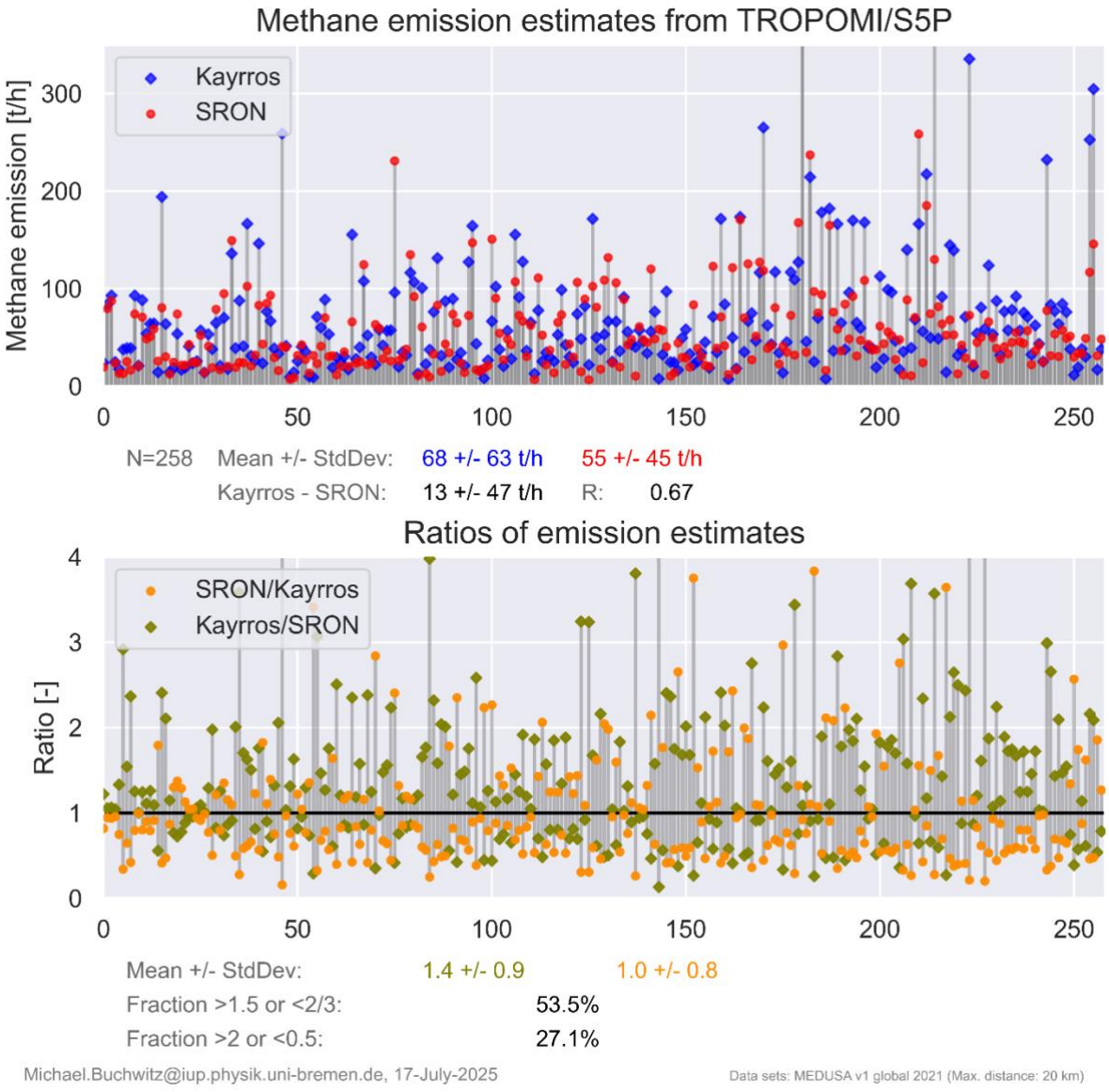


Figure 4.2-17: *Comparison of the collocated Kayrros and SRON methane emission estimates obtained from the year 2021 global data sets.Top: Emission estimates of Kayrros (blue symbols) and SRON (red). Bottom: Ratios of emissions: SRON/Kayrros (orange) and Kayrros/SRON (olive green).*

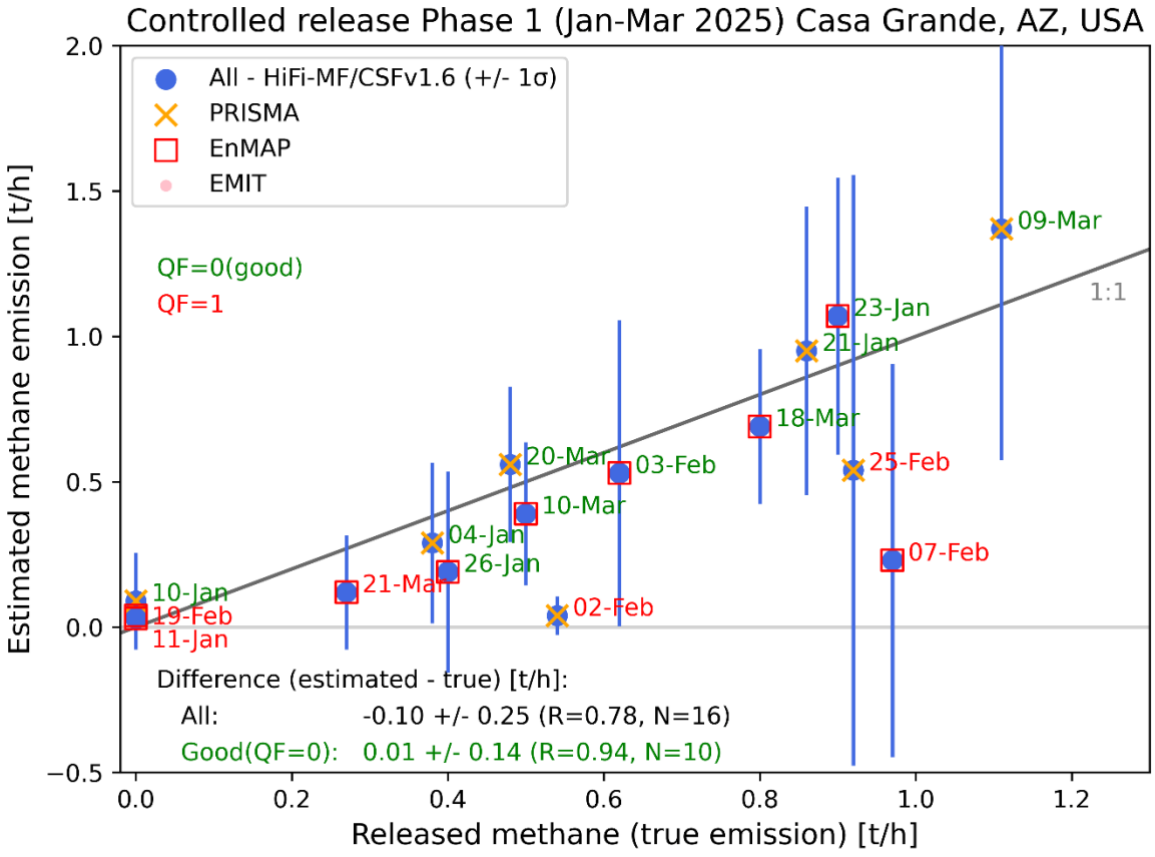


Figure 4.2-18: *Results from single-blind Phase 1 of the controlled methane release experiment as conducted in Casa Grande, Arizona, USA, by teams from Stanford University and University of Michigan during January – March 2025. Shown are comparisons of satellite-derived estimated methane emissions as obtained using the IUP-UB CSF v1.6 algorithm to PRISMA and EnMAP XCH_4 enhancements compared to released ("true") emissions as made available ("unblinded") by the Stanford / Michigan teams to all satellite retrieval teams after they have submitted their estimates. Source: Buchwitz et al., 2025.*

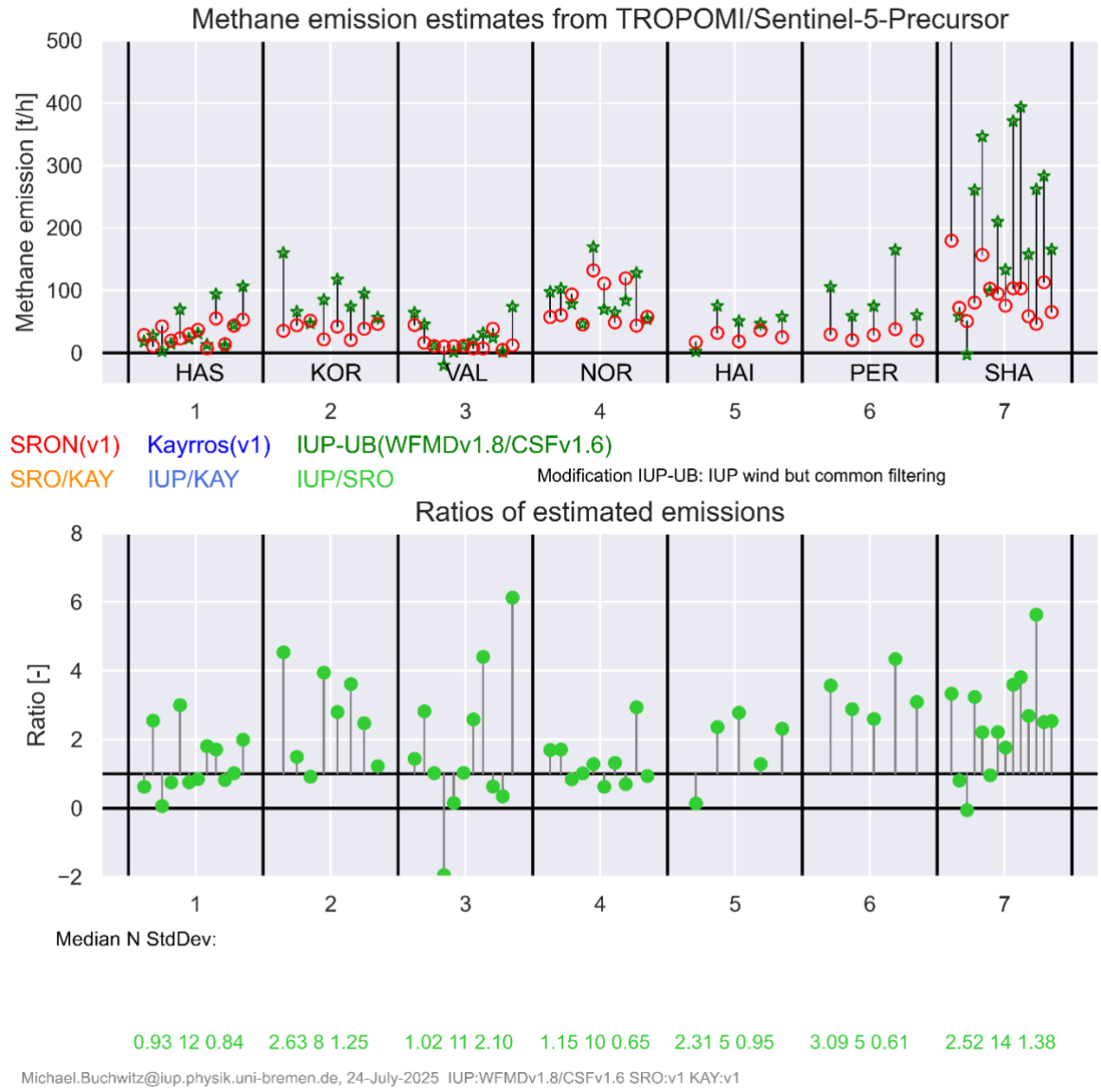


Figure 4.2-19: *Similar as* ***Figure 4.2-12*** *but limited to the IUP-UB versus SRON comparison. The mean value of the ratios is 2.0 ± 1.4 (R = 0.76, N = 65).*

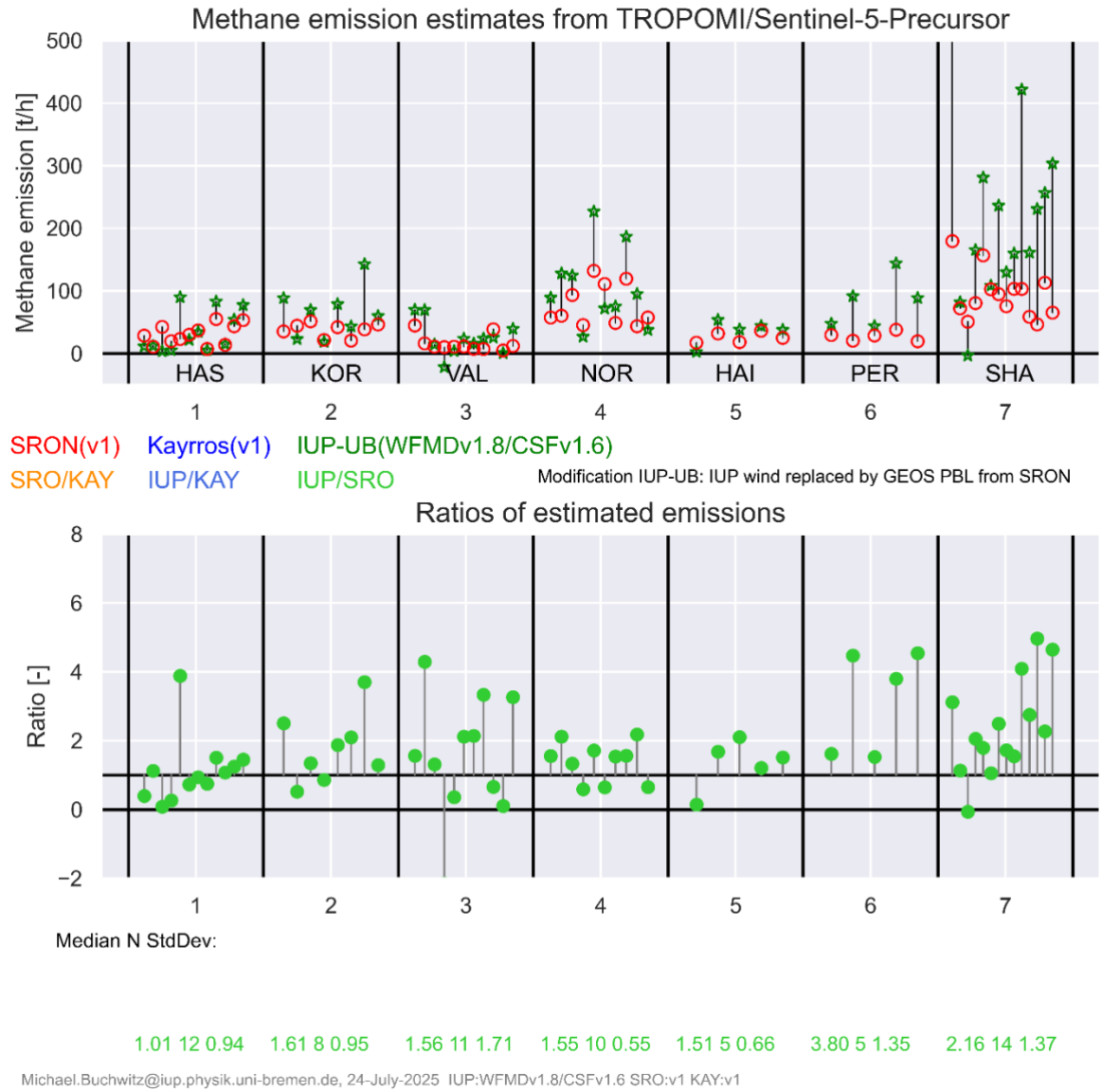


Figure 4.2-20: *Similar as* ***Figure 4.2-19*** *but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by GEOS PBL wind speed. The mean value of the ratios is 1.8 ± 1.3 (R = 0.79, N = 65).*

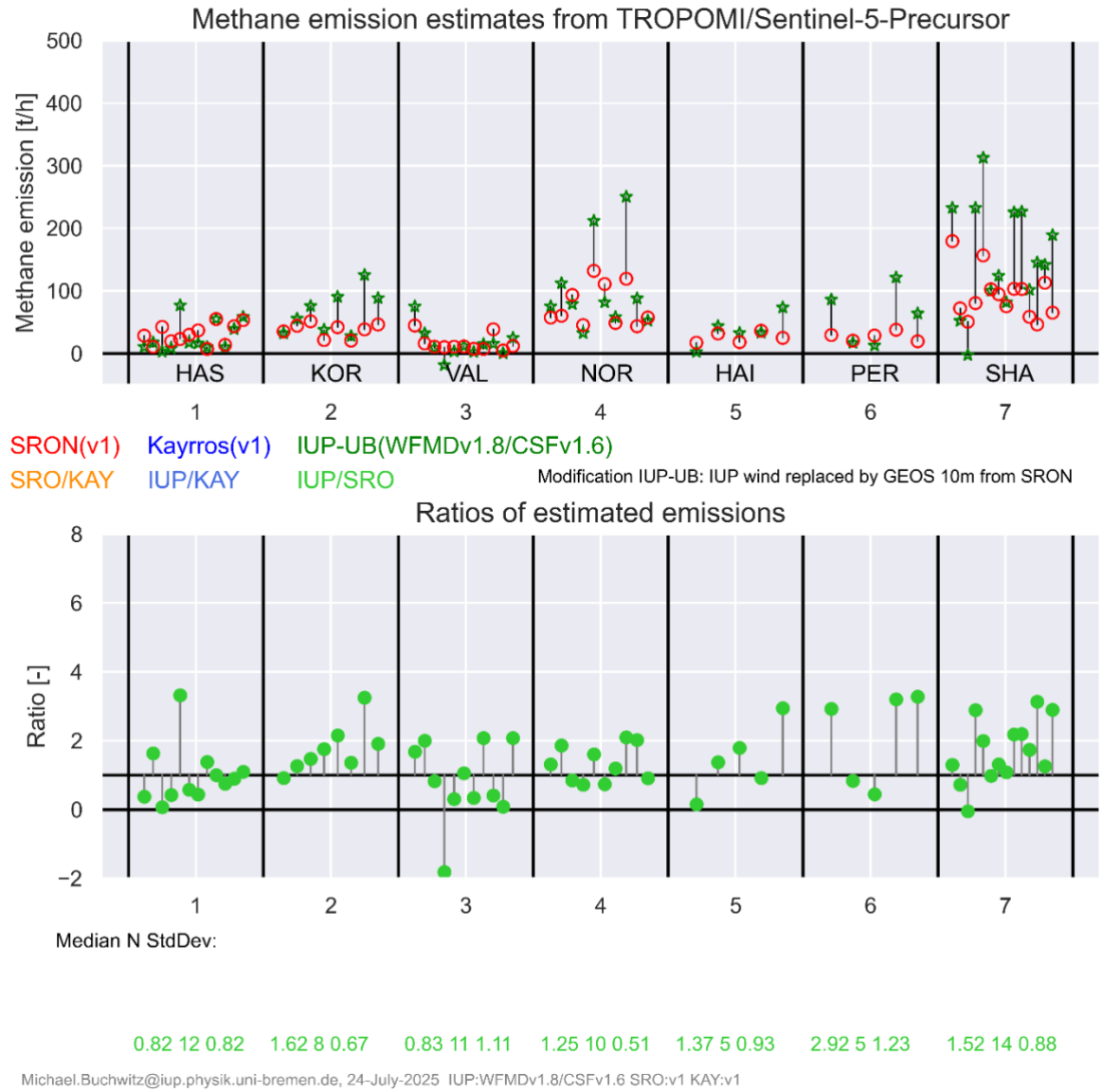


Figure 4.2-21: *Similar as* ***Figure 4.2-19*** *but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by GEOS 10m wind speed. The mean value of the ratios is 1.4 ± 1.0 (R = 0.83, N = 65).*

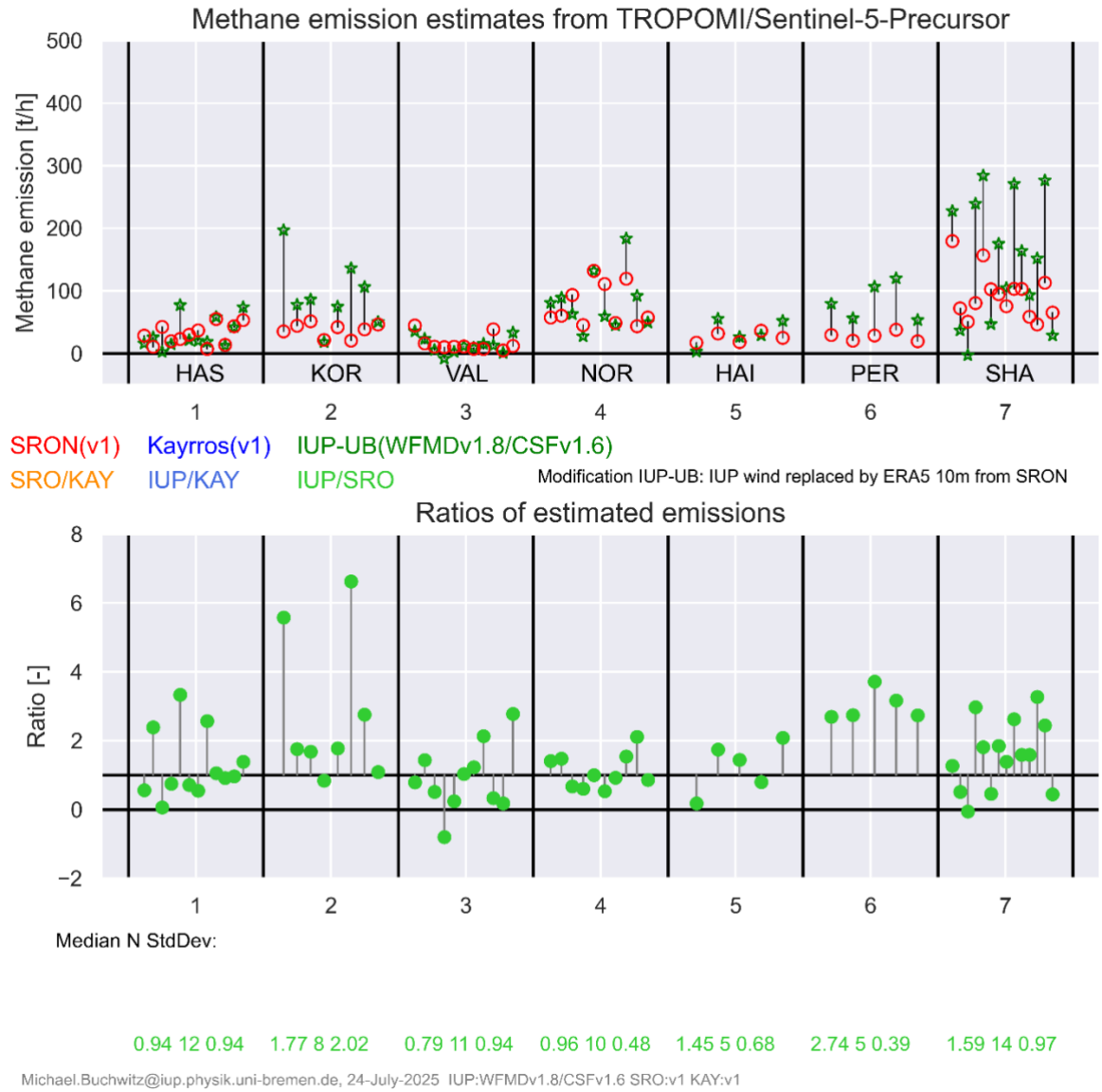


Figure 4.2-22: *Similar as* ***Figure 4.2-19*** *but with modified IUP-UB emission estimates obtained by replacing ERA5 PBL average wind speed by ERA5 10m wind speed. The mean value of the ratios is 1.6 ± 1.3 (R = 0.72, N = 65).*

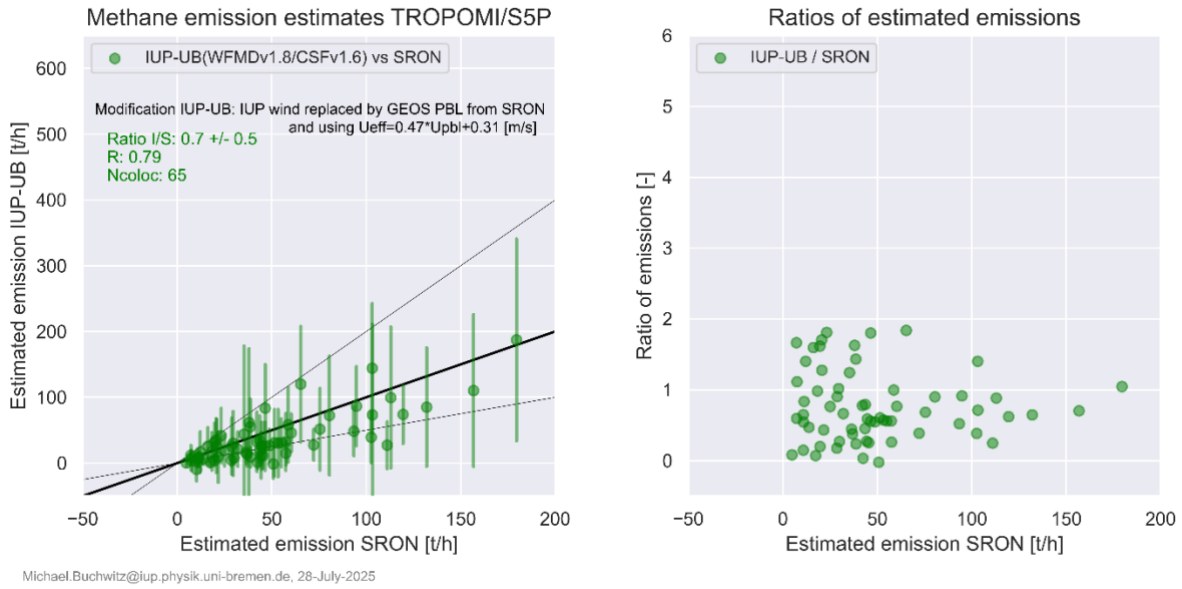


Figure 4.2-23: *Similar as* ***Figure 4.2-14*** *but limited to the IUP-UB versus SRON comparison and using for IUP-UB an effective wind speed Ueff computed using the "calibrated" GEOS PBL wind speed also used by SRON for their IME algorithm (i.e., using Ueff = 0.47*U_{PBL}+0.31).*

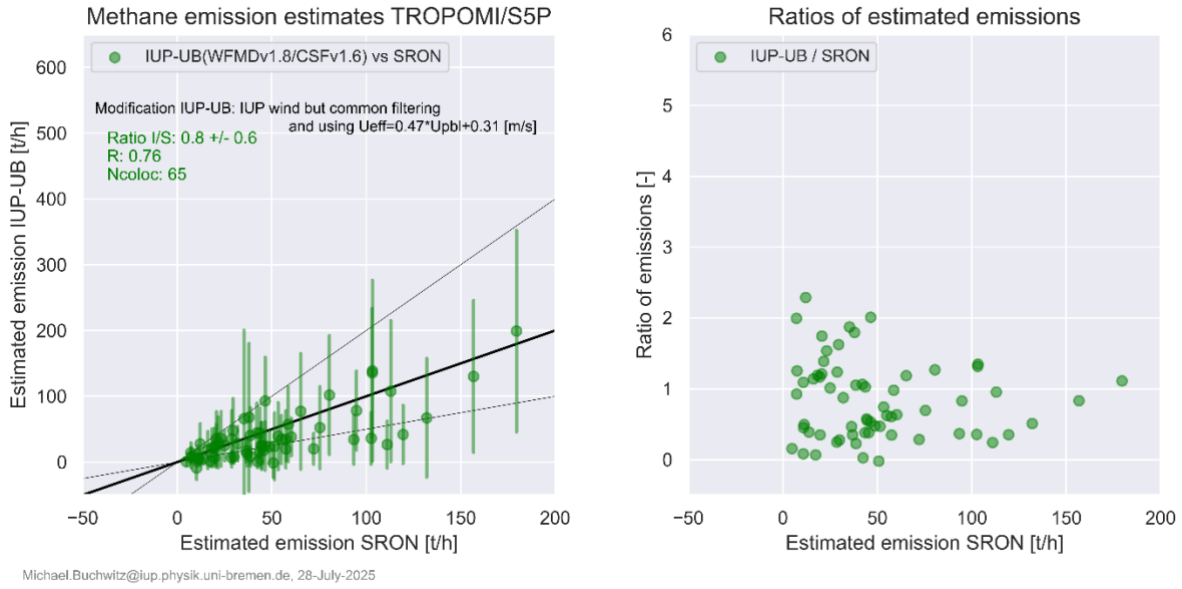
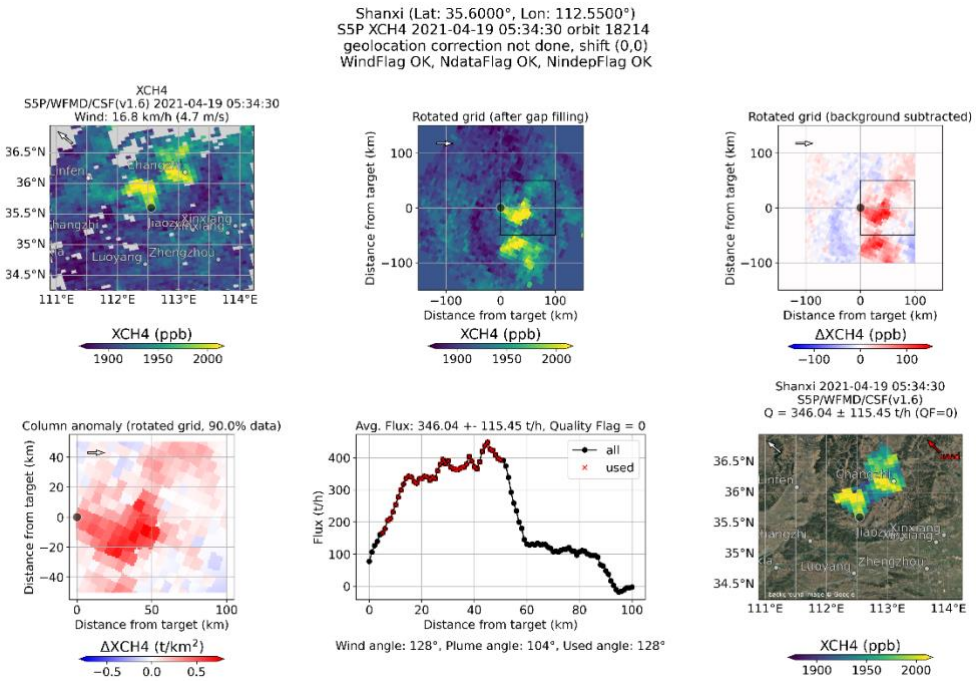


Figure 4.2-24: *Similar as* ***Figure 4.2-23*** *but using for IUP-UB a "calibrated" ERA5 PBL wind speed computed using SRON's Ueff formula (see* ***Figure 4.2-23****).*

(a)



(b)

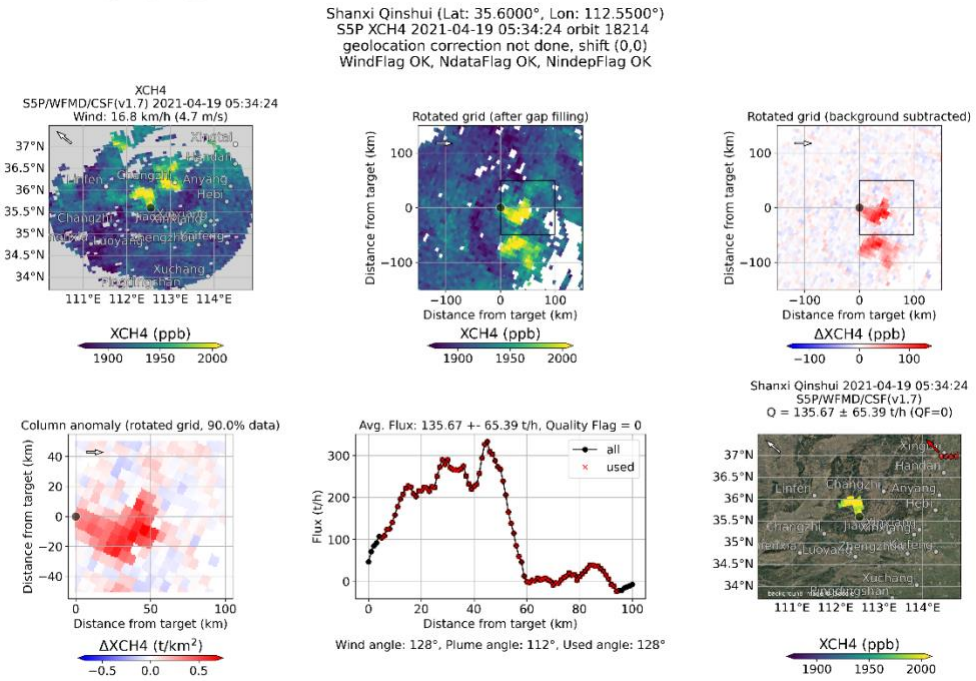
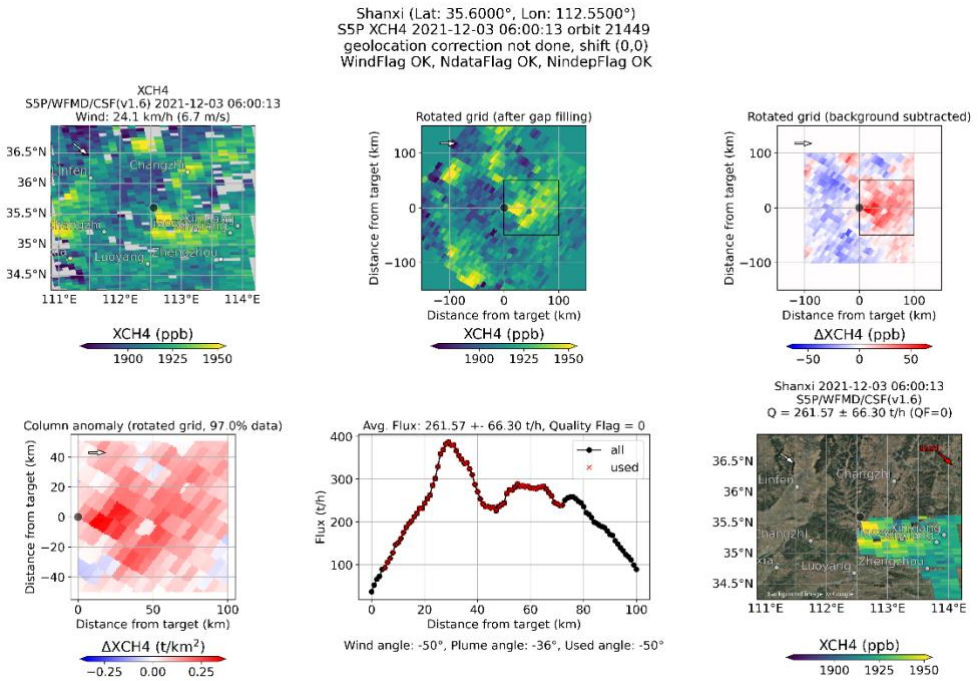


Figure 4.2-25: *IUP-UB TROPOMI CSF emission estimates for target region SHA (Qinshui, Shanxi, China) on 19-April-2021. (a): CSF v1.6. (b): CSF v1.7.*

(a)



(b)

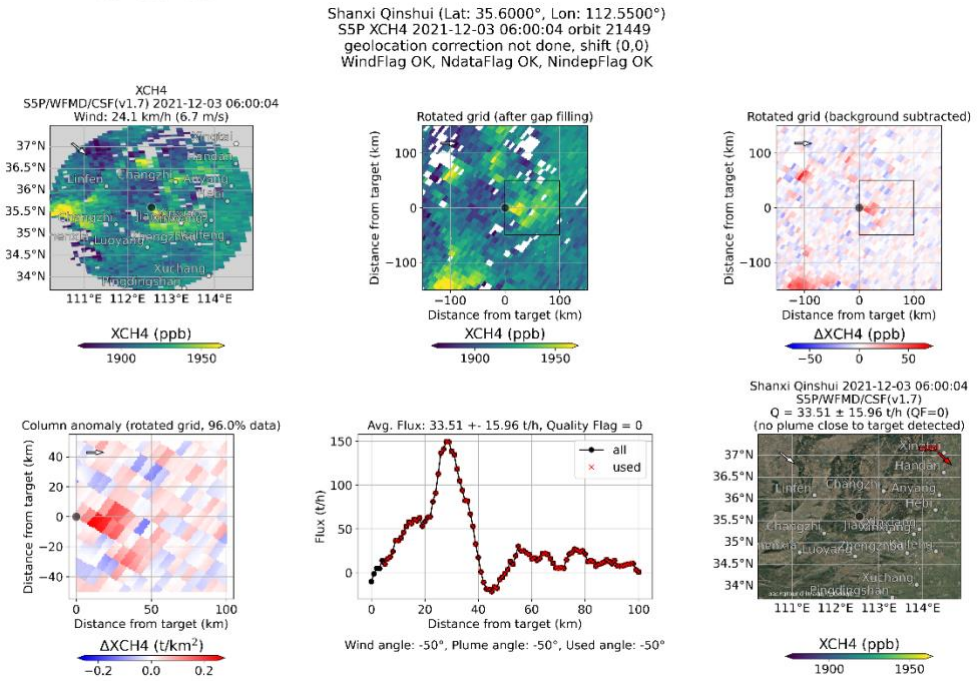
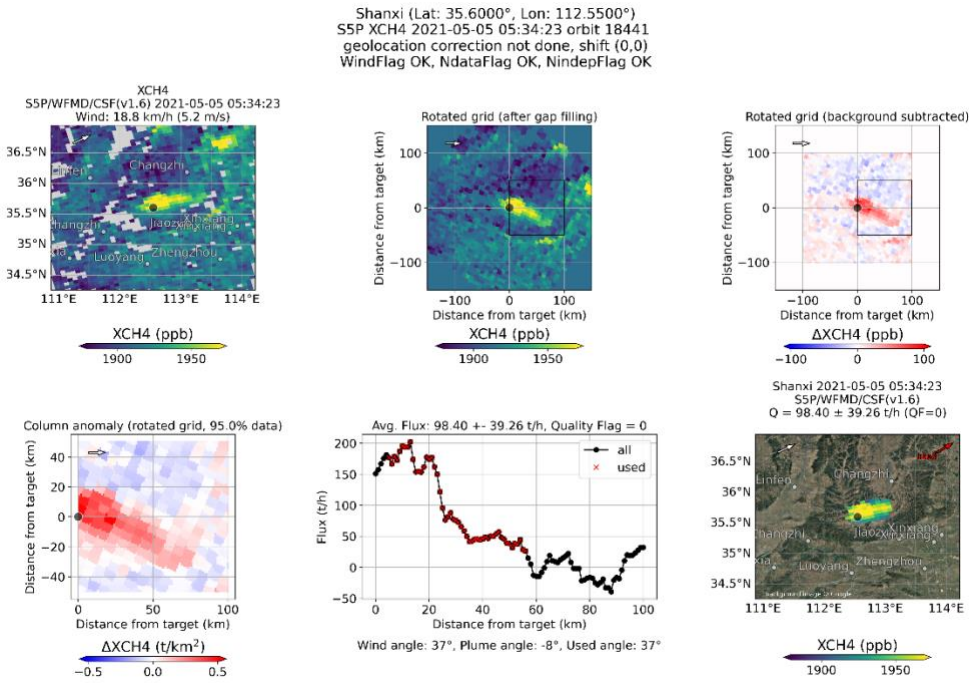


Figure 4.2-26: *IUP-UB TROPOMI CSF emission estimates for target region SHA (Qinshui, Shanxi, China) on 3-December-2021. (a): CSF v1.6. (b): CSF v1.7.*

(a)



(b)

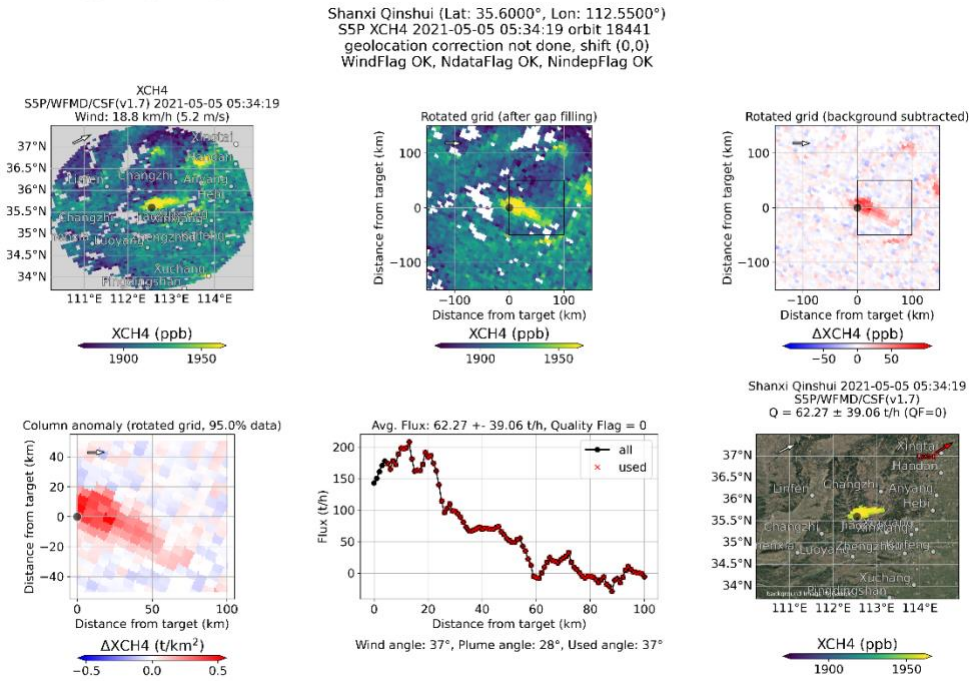


Figure 4.2-27: *IUP-UB TROPOMI CSF emission estimates for target region SHA (Qinshui, Shanxi, China) on 5-May-2021. (a): CSF v1.6. (b): CSF v1.7.*

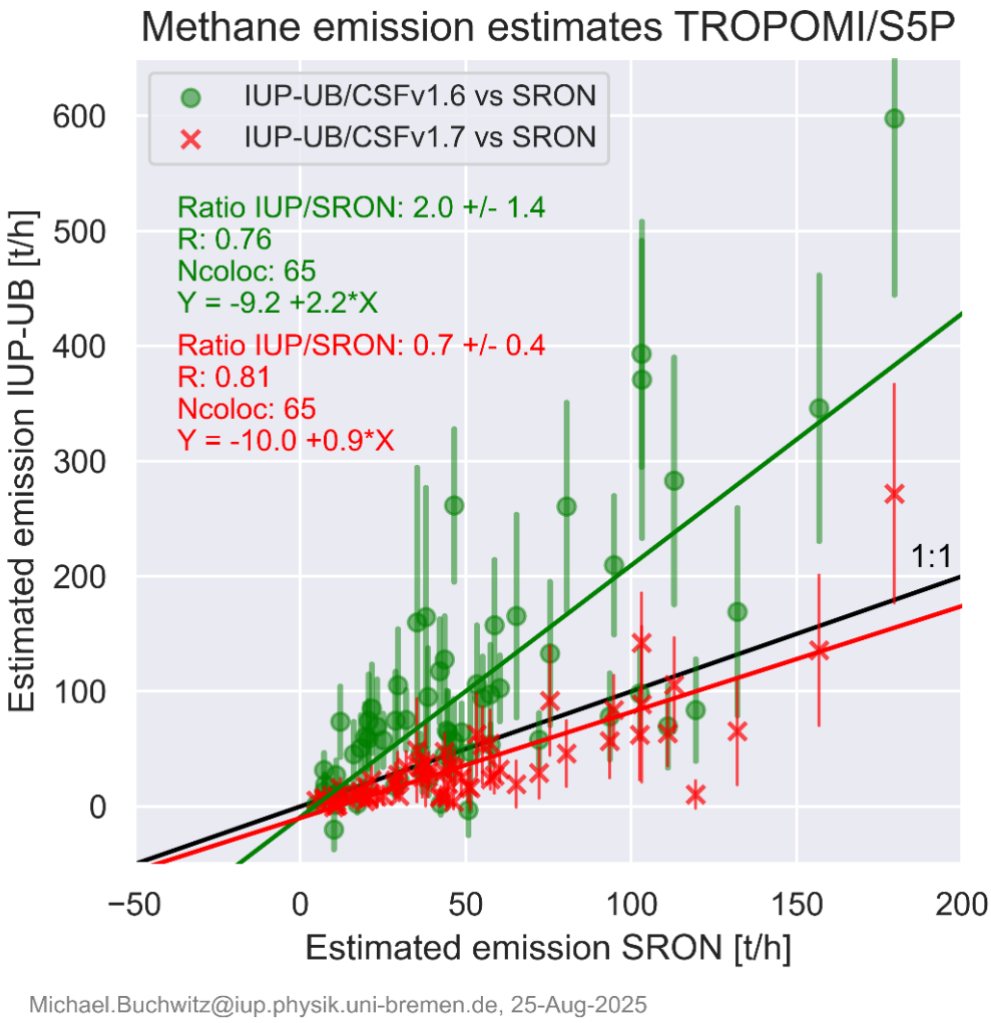


Figure 4.2-28: *Similar as* ***Figure 4.2-14 (left)*** *but limited to the IUP-UB and SRON comparison and with IUP-UB CSF v1.7 results added (in red).*

4.3. Summary of comparisons at target locations

Detailed comparisons of methane emission estimates (Level 4 products) at 10 target locations have been presented in this section as obtained from TROPOMI satellite XCH_4 retrievals (Level 2 products) using the 3 different Level 2 to Level 4 algorithms as developed and used by SRON, Kayrros and IUP-UB. The comparison shows that, as expected, the IUP-UB Cross-Sectional-Flux (CSF) algorithm provides typically the largest number of methane emission estimates for a given target location. This is because this algorithm is not based on emission plume detection in contrast to the other 2 algorithms. Instead, only a small set of selection criteria is used in order to decide if the satellite data can be used for emission estimation or not: Data coverage must be dense enough (e.g., > 90% of the area used for the CSF analysis needs to be covered by "good" (e.g., cloud free) data independent of the spatial pattern of the methane concentration field) and wind speed should not be too low (> 1 m/s). If these criteria are met than the quality flag (QF) attached to the corresponding emission estimate is set to "good" (QF=0) (otherwise it is set to 1). Consequently, estimated emissions using IUP-UB's CSF algorithm may also be zero or even negative in case no methane plume is present and because XCH_4 retrievals are noisy. The smallest number of emission estimates are provided by the Kayrros algorithm due to a conservative detection method and their quantification approach, which is based on detailed modelling, which is challenging due to the required good knowledge of meteorological information (e.g., wind field), issues related to turbulence, etc. The number of estimates provided by SRON is in between the number of estimates of Kayrros and IUP-UB. For a detailed comparison of the global SRON and Kayrros emission estimates see **Sect. 5**, where it is also shown that the SRON algorithm produces a significantly larger number of emission estimates compared to the Kayrros algorithm and additional explanations are given why this is the case.

It needs to be pointed out that the IUP-UB CSF algorithm as used for MEDUSA is new and still under development. Most of the results generated by IUP-UB for MEDUSA have been produced with version 1.6 of this algorithm. CSF v1.6 is also the algorithm used for emission estimates from the Hyperspectral Imagers (HI) EnMAP, PRISMA and EMIT (see **Annex B** to document AIR). Results reported in the following refer to comparisons with v1.6 except if stated otherwise.

Primarily due to the low number of Kayrros estimates and the fact that often only one algorithm provided an estimate for a given target location and satellite overpass, the number of collocations is small. At 3 of the 10 target locations no collocations have been found. For the 7 other locations where collocations exist, on average good to reasonable agreement between the IUP-UB and the SRON estimates has been found for 3 of these targets but for the 4 other targets, the IUP-UB estimates were significantly higher (on average by more than a factor of 2). In contrast, the agreement between Kayrros and SRON was found to be much better with agreement typically within some 10%. But this finding is based on only a very small number of collocations and more meaningful are the global comparisons results reported separately.

Several assessments have been carried out to explain the relative overestimation of the IUP-UB CSF v1.6 emission estimates compared to SRON at some of the target locations. It has been identified that the different Level 2 input data are not the driver for these differences, and it is argued that also the wind speed used by IUP-UB is unlikely the reason for a factor of 2 or larger differences. IUP-UB uses for TROPOMI the Planetary Boundary Layer (PBL) average of the wind speed, which is in line with Varon et al., 2019, who found that the PBL average wind speed is appropriate for TROPOMI CSF-based emission estimates.

One must note that the target region comparisons are based on only a small number of targets and with only a small number of collocations per target. This raises the question how representative the target region comparisons are. Additional comparisons for several other locations have been conducted independently by B. Dils (BIRA-IASB) in the context of MEDUSA work package (WP) 220. These comparisons show on average very good agreement between the emission estimates as obtained by IUP-UB with CSF v1.6, SRON and Kayrros. Here the main findings based on correlation (scatter) plots are: slope $\beta = 1.01$ ($N = 26$, $R^2 = 0.74$) for SRON vs IUP-UB, $\beta = 0.92$ ($N = 22$, $R^2 = 0.69$) for Kayrros vs IUP-UB and $\beta = 1.05$ ($N = 22$, $R^2 = 0.55$) for SRON vs Kayrros. This indicates that all three algorithms result in very similar emission estimates on average (as all slopes are close to 1) although (as indicated by the moderate R^2 values) the individual emission estimates may vary significantly (by often several 10%). This indicates that the level of agreement or disagreement of the results obtained with the different algorithms depends on which subset is compared.

Nevertheless, additional sensitivity studies have been carried out to find out if the identified differences of the IUP-UB CSF v1.6 results at the target regions relative to SRON can be explained. These activities resulted in a new version 1.7 of the IUP-UB CSF algorithm. The main differences of v1.7 compared to v1.6 are a modified background correction and a waiver of the 3-hour plume length limit used for v1.6. It is shown that v1.7 results are in much better agreement with the SRON estimates compared to v1.6 at the selected target regions indicating that the addressed aspects may explain at least parts of the observed differences.

So far only a limited amount of data has been processed and analyzed with CSF v1.7 and further studies are needed to find out if v1.7 is better than v1.6 also for cases not yet investigated.

Due to the limited number of collocations with the Kayrros estimates at the selected 10 target regions, the target region comparison focused on comparisons between the IUP-UB and SRON algorithms. For a more meaningful comparison of the SRON and the Kayrros algorithms, please see the following section focusing on global comparisons.

5. Global comparisons

In this section the global intercomparison results are presented. The target region comparison results have been shown separately in **Sect. 4**.

The global intercomparison is based on 2 data sets, namely the data sets of estimated methane emissions as derived from TROPOMI/S5P using the algorithms from SRON and Kayrros (for details see **Sect. 3.1**).

5.1. Global intercomparison data, methods and results

This section presents the two datasets in more detail, then describes how plumes from the two datasets are matched up. Match-up results are then detailed over three subsections corresponding to the three main cases of overlap: one-to-one matching plumes and plumes detected by one group but not reported by the other, for both groups.

5.1.1. SRON and Kayrros global TROPOMI plume detection datasets

SRON and Kayrros both provided a global dataset of all methane plumes detected in TROPOMI data for the year 2021. **Figure 5.1.1-1** shows the binned spatial distributions of these two datasets, and their (binary) difference. Kayrros' dataset contains 1060 plumes, and SRON's 3233. Both datasets show detected methane emission plumes in prominent oil & gas and coal mining areas around the world, e.g. the Permian basin in the USA, Hassi Messaoud region in Algeria, along Russian pipelines, the West coast of Turkmenistan, the Shanxi coal mining region in China, and along the East Coast of Australia. Both datasets also find plumes arising from urban areas, like Buenos Aires in Argentina, Madrid in Spain, and Delhi in India. **Figure 5.1.1-1** also provides the binary difference between the gridded maps of both sets (lower right panel), showing where one set includes plumes and the other does not. In general, there are many more locations where SRON detects plumes and Kayrros does not, than the opposite. These are overall located in the vicinities of the previously cited regions, but these also include separate locations as well as, for example, the Casablanca (Morocco), Istanbul (Turkey), and Sao Paulo (Brazil) city areas, coal mines in Poland and oil & gas production sites in Yemen.

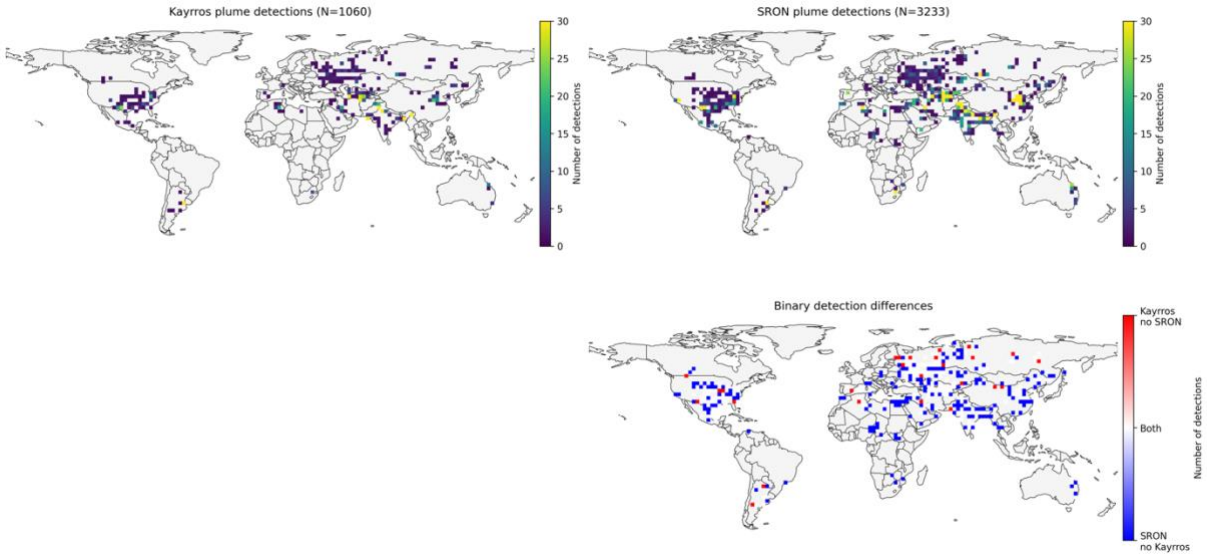


Figure 5.1.1-1: Spatial gridded (2.5° x 2.5°) distribution maps of methane plumes detected in TROPOMI data for the year 2021 by Kayrros (top left), SRON (top right) and their binary difference (bottom right), highlighting where at least one plume has been detected by one group only.

5.1.2. Plume match-making approach for SRON and Kayrros datasets

SRON and Kayrros global TROPOMI methane plume datasets are matched up by examining orbit- and pixel-wise plume mask (non-)overlaps between both sets. We define a match when two plume masks of different sets share at least one pixel within the data of a same orbit. Three different outcomes are possible, illustrated in **Figure 5.1.2-1**: (1) one unique SRON-detected plume overlaps one unique Kayrros-detected plume (a, in **Figure 5.1.2-1**); (2) SRON and Kayrros plumes are non-overlapping (b, in **Figure 5.1.2-1**); and (3) several plumes from one group overlap with a unique plume from the other (c and d, in **Figure 5.1.2-1**, with two Kayrros plumes overlapping one SRON plume).

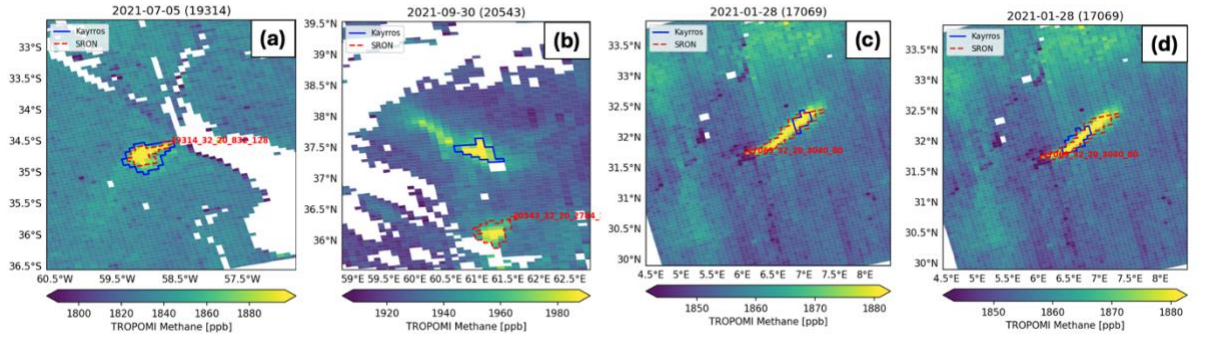


Figure 5.1.2-1*: Three different types of orbit- and pixel-wise plume matching results between SRON and Kayrros datasets: a unique plume matches a unique plume (a); plumes do not*

overlap (b); and one plume matches several plumes (c and d). The titles of the frames show orbit dates and numbers.

The plume matching results are provided in **Table 5.1.2-1**. We find very few one-to-multiple matches compared to other categories. Consequently, we focus our analysis on the 643 SRON-Kayrros one-to-one matches, and on the 386 and 2568 Kayrros-only and SRON-only reported plumes, respectively. The following subsections analyze relevant comparison questions for these three sets.

	SRON	**Kayrros**
One-to-one matches	643	643
one-to-multiple matches	*22*	*31*
non-matching	2568	386
TOTAL	3233	1060

***Table 5.1.2-1**: SRON and Kayrros plume match-making results.*

5.1.3. Results: one-to-one SRON – Kayrros plume matches

For one-to-one matching plumes, we study the difference between Kayrros and SRON emission rate quantifications obtained for each matching plume.

Kayrros quantifications rely on the HYSPLIT atmospheric transport model that is run for a baseline emission duration of 7 hours and then compared to the observed plume. If simulated and observed plume extents do not match, simulations for different emission durations are run to see if a match with an alternate duration can be found. The agreement between observed and simulated plumes is visually evaluated, and only the best agreeing cases are quantified. Consequently, a detected plume in Kayrros dataset may not be quantified: among the 643 one-to-one matching plumes, 306 are quantified. Thus, our emission rate comparison is based on 306 matching plumes.

While examining Kayrros plume masks, we identified that some of them include pixels with missing TROPOMI data (see **Figure 5.1.3-1**). This is explained by Kayrros filling pixels with missing data by employing biharmonic inpainting to be able to run their computer-vision approach used to detect plumes in TROPOMI observations. We found that the presence of (inpainted) missing data pixels in Kayrros plume mask is an important aspect driving part of the difference between SRON and Kayrros emission quantifications, even if inpainted pixels are not used in the quantification step.

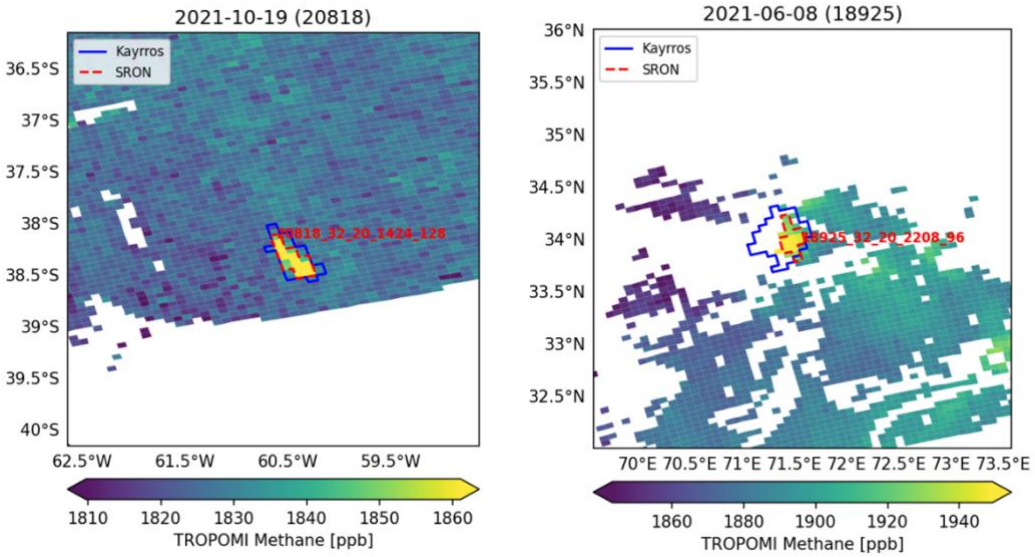


***Figure 5.1.3-1**: Examples of two Kayrros plumes (blue), one where the Kayrros plume mask is full of TROPOMI pixels with valid data (left), and another one where the Kayrros plume mask includes TROPOMI pixels with missing data (right).*

Regarding SRON emission quantifications, all plumes are quantified using the Integrated Mass Enhancement (IME) method in an ensemble approach that includes ensemble members with three different wind speed products: ERA5 and GEOS-FP 10m wind speed and Planetary Boundary Layer-averaged wind speed from GEOS-FP (see **Sect. 3.1.1**). We find that the actual wind-speed used in SRON quantification is a relevant aspect to discuss in the comparison of one-to-one matches emission rates.

Figure 5.1.3-2 provides Kayrros – SRON emission rate comparison results for the different subsets and cases discussed above: separating SRON results depending on the used wind speed (different lines) and separating plumes depending on whether Kayrros plume masks include TROPOMI pixels with missing data or not (different columns). The results are also summarized in **Table 5.1.3-1**.

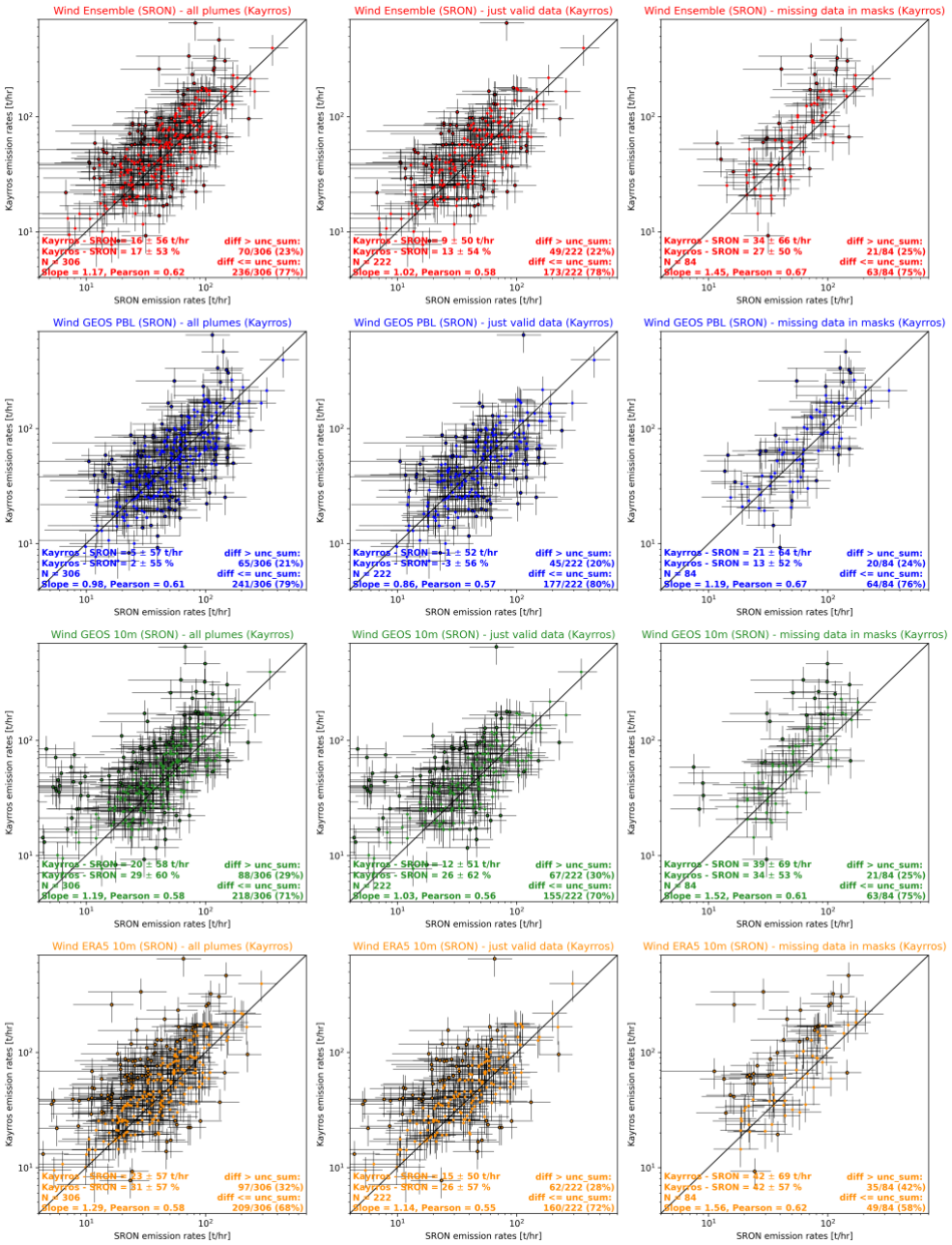


***Figure 5.1.3-2**: Comparison of Kayrros and SRON emission rates for all one-to-one matches and quantified plumes (left, N=306), for all plumes with only valid data in their masks (center, N=222) and for plumes with missing pixels in Kayrros plume masks (right, N=84), using the full SRON wind ensemble result (top line, red) and using just GEOS PBL averaged (2nd line, blue), GEOS-FP 10 m (3rd line, green) and ERA5 10 m (bottom, orange) wind speeds. One-to-one matching plumes highlighted with black outlines do not have overlapping uncertainty ranges.*

	All plumes (N=306)	**Plumes with only valid pixels (N=222)**	**Plumes that include pixels with missing data (N=84)**
Full SRON wind speed ensemble	16 ± 56 t/hr 17 ± 53 % (s = 1.17, r = 0.62) n = 70/306 (23%) y = 236/306 (77%)	9 ± 50 t/hr 13 ± 54 % (s = 1.02, r = 0.58) n = 49/222 (22%) y = 173/222 (78%)	34 ± 66 t/hr 27 ± 50 % (s = 1.45, r = 0.67) n = 21/84 (25%) y = 63/84 (75%)
SRON only using GEOS-PBL	5 ± 57 t/hr 2 ± 55 % (s = 0.98, r = 0.61) n = 65/306 (21%) y = 241/306 (79%)	-1 ± 52 t/hr -3 ± 56 % (s = 0.86, r = 0.57) n = 45/222 (20%) y = 177/222 (80%)	21 ± 64 t/hr 13 ± 52 % (s = 1.19, r = 0.67) n = 20/84 (24%) y = 64/84 (76%)
SRON only using GEOS-FP 10m	20 ± 58 t/hr 29 ± 60 % (s = 1.19, r = 0.58) n = 88/306 (29%) y = 218/306 (71%)	12 ± 51 t/hr 26 ± 62 % (s = 1.03, r = 0.56) n = 67/222 (30%) y = 155/222 (70%)	39 ± 69 t/hr 34 ± 53 % (s = 1.52, r = 0.61) n = 21/84 (25%) y = 63/84 (75%)
SRON only using ERA5 10 m	23 ± 57 t/hr 31 ± 57 % (s = 1.29, r = 0.58) n = 97/306 (32%) y = 209/306 (68%)	15 ± 50 t/hr 26 ± 57 % (s = 1.14, r = 0.55) n = 62/222 (28%) y = 160/222 (72%)	42 ± 69 t/hr 42 ± 57 % (s = 1.56, r = 0.62) n = 35/84 (42%) y = 49/84 (58%)

***Table 5.1.3-1**: Kayrros – SRON difference results (absolute and relative to the mean) for the cases explored in **Figure 5.1.3-2**, with s, the Ordinary Least Squares fitted Kayrros/SRON slope and r, the Pearson correlation coefficient, n, the fraction of plumes where Kayrros and SRON uncertainty ranges do not overlap, and y, the fraction where they overlap.*

First, considering all the 306 quantified one-to-one matches as delivered by Kayrros and SRON (top left in **Figure 5.1.3-2**), we find that Kayrros reports slightly higher estimates than SRON, with an (insignificant) difference of 16 ± 56 t/hr. The correlation between Kayrros and SRON quantifications is high, with a Pearson correlation coefficient of 0.62. This result is especially remarkable considering that SRON and Kayrros are using (1) completely different quantification approaches with a mass-balance approach and atmospheric simulation fits; and (2) completely different meteorological products, with GEOS-FP and ERA5 for SRON and GFS from NCEP for Kayrros. This consistency – on average – between Kayrros and SRON despite completely different methods and input meteorological data builds confidence in the obtained emission rate quantifications. However, it is also important to note that, while differences are small on average, the scatter of Kayrros – SRON difference is quite significant with 53% of relative (to the average) difference scatter, which translates into a quarter of quantifications disagreeing by a factor 2 or more. This scatter can also be examined considering the uncertainties reported by Kayrros and SRON. On average, their emission rates have 30% and 41% relative uncertainties, respectively.

Consequently, 77% of Kayrros and SRON quantifications agree within their reported uncertainty ranges, and 23% do not. The following efforts aim at trying to identify possible explanation elements of this significant scatter.

Figure 5.1.3-2 also explores the contribution of Kayrros including missing pixels in their plume masks to explain the Kayrros – SRON difference results. First, considering Kayrros plumes that include TROPOMI pixels with missing data increases the difference and its scatter: for full SRON wind ensemble results, the difference increases from 9 ± 50 t/hr to 34 ± 66 t/hr using plumes with only valid data, and using only plumes with missing data, respectively. The baseline comparison result of Kayrros and SRON quantifications is a mix of these two extreme cases, with a difference of 16 ± 56 t/hr. To exclude the contribution of Kayrros sometimes including TROPOMI pixels with missing data in their masks, we focus on the 222 plumes that only include valid pixels in Kayrros masks for the rest of this analysis.

Figure 5.1.3-2 finally also explores the contribution of the wind speed used in the SRON quantification to explain the Kayrros – SRON difference results. Considering all the plumes with only valid data in Kayrros masks, the SRON results only based on 10 m wind speed (GEOS-FP and ERA5) are slightly lower than the ones obtained with GEOS PBL-averaged wind speeds. This shows a slight inconsistency in SRON effective wind speed calibrations for 10-m and PBL-average wind speeds; efforts are currently underway at SRON to improve the effective wind speed calibrations that are used. However, results using the full SRON ensemble results show the highest correlation against Kayrros compared to results obtained with separate wind speed origins. This highlights the complementarity of the different winds included in the SRON ensemble, and the importance/value of employing such an approach.

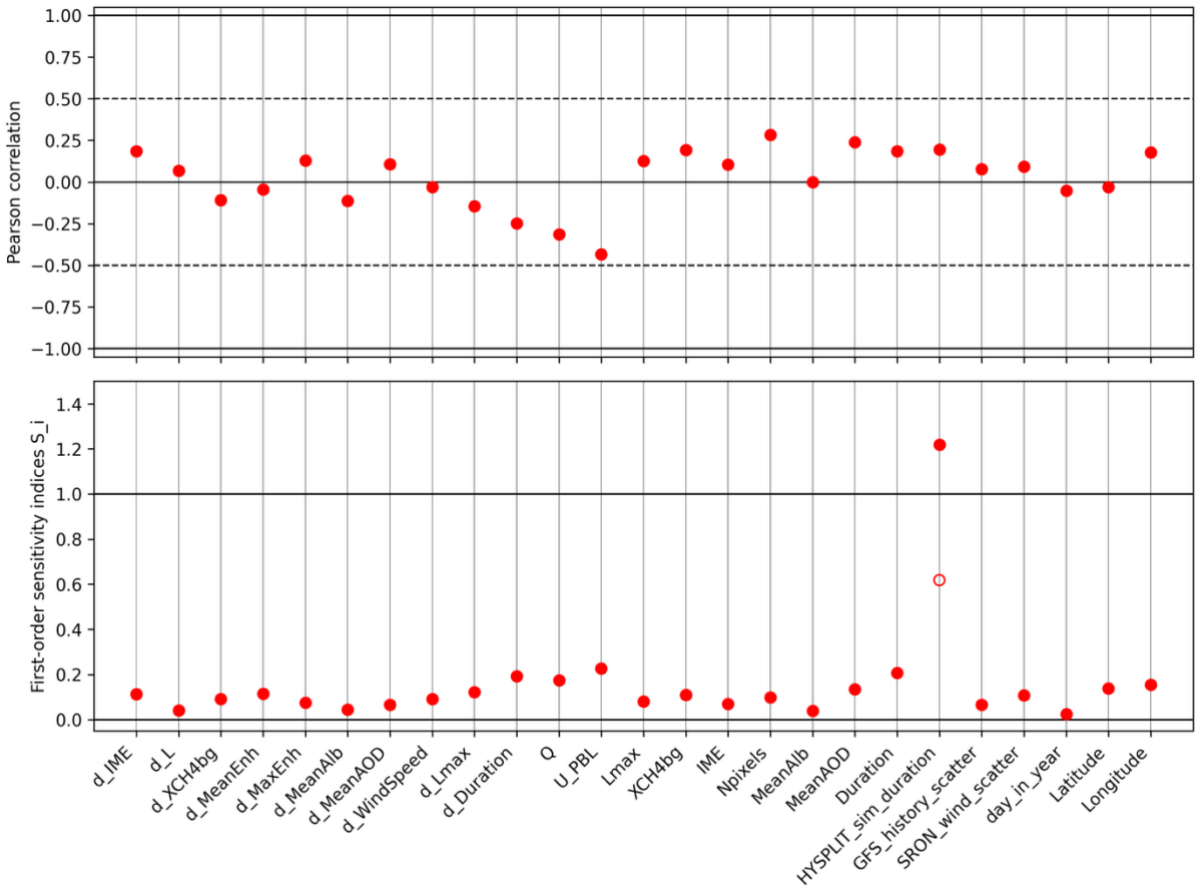


***Figure 5.1.3-3**: Pearson correlation coefficient between the Kayrros – SRON emission rate difference and different plume features for one-to-one matching plumes (top) and First-order sensitivity index obtained for these features (bottom). The physical definition of all these features is provided in **Table 5.1.3-2**. The empty circle for HYSPLIT_sim_duration is a special filter case, see main text.*

We now seek to explore explanation elements of the significant scatter shown by the SRON – Kayrros difference. We focus on the SRON wind ensemble results for the 222 plumes whose Kayrros plume masks only include valid TROPOMI pixels, and we exclude one outlier with a dQ (Kayrros – SRON) above 400 t/hr (see **Figure 5.1.3-2**). Consequently, we perform this analysis on 221 plumes. We explore many different features that may or may not influence the estimated emission rate and/or its difference. **Figure 5.1.3-3** provides the Pearson correlation coefficient between the Kayrros – SRON emission difference and each explored feature, and the First-order sensitivity index (Lo Piano et al., 2021) showing which fraction of the Kayrros – SRON emission difference variance is explained by each explored feature. Overall, we find that all features show relatively weak correlations with the Kayrros – SRON emission rate differences, and that they only explain a small fraction of the difference variance, except for the Kayrros HYSPLIT simulation duration, showing a contribution to the difference variance larger than 1. This non-mathematical result is explained by Kayrros HYSPLIT simulation durations only taking a few discrete values (2, 3, 4, 5, 6, 7 and 10

hours), and that there are only single plumes with simulation durations of 6 and 10 hours. Removing these two plumes from the analysis yields a relatively high First-order sensitivity index for the Kayrros HYSPLIT simulation duration (~0.6), but now lower than 1. We conclude from **Figure 5.1.3-3** that the only feature worth exploring is the Kayrros HYSPLIT simulation duration.

Feature name in Figure 5.1.3-3	**Definition**
d_IME	Kayrros – SRON difference in integrated mass enhancement [kg]
d_L	Kayrros – SRON difference in plume length (defined as the square root of the plume mask area) [m]
d_XCH4bg	Kayrros – SRON difference in background XCH4 values [ppb]
d_MeanEnh	Kayrros – SRON difference in mean enhancement value in the plume mask [ppb]
d_MaxEnh	Kayrros – SRON difference in maximum enhancement value in the plume mask [ppb]
d_MeanAlb	Kayrros – SRON difference in mean albedo value in the plume mask [-]
d_MeanAOD	Kayrros – SRON difference in mean aerosol optical depth in SWIR value in the plume mask [-]
d_WindSpeed	Kayrros – SRON difference in PBL-averaged wind speed [m/s]
d_Lmax	Kayrros – SRON difference in maximum length in plume mask [m]
d_Duration	Kayrros – SRON difference in transport duration along the maximum plume length [hr]
Q	SRON emission rate [t/hr]
U_PBL	GEOS-FP PBL-averaged wind speed [m/s]
Lmax	Maximum geometric length in the SRON plume mask [m]
XCH4bg	XCH4 background value computed by SRON [ppb]
IME	Integrated Mass Enhancement in SRON plume masks [kg]
Npixels	Number of pixels in Kayrros plume masks
MeanAlb	Average albedo in SRON plume masks
MeanAOD	Average aerosol optical depth in SRON plume masks
Duration	Transport duration in the SRON plume mask, based on L_max/U_PBL [hr]
HYSPLIT_sim_duration	HYSPLIT simulation emission duration used by Kayrros to fit the detected plume [hr]
GFS_history_scatter	Standard deviation of the GFS wind speed over the time range of the simulated [m/s]

SRON_wind_scatter	Maximum range the SRON wind speed ensemble [m/s]
day_in_year	Number of days since 2021-01-01T00:00:00
Latitude	Latitude of the plume source [°]
Longitude	Longitude of the plume source [°]

***Table 5.1.3-2**: Definition of the features used in **Figure 5.1.3-3**.*

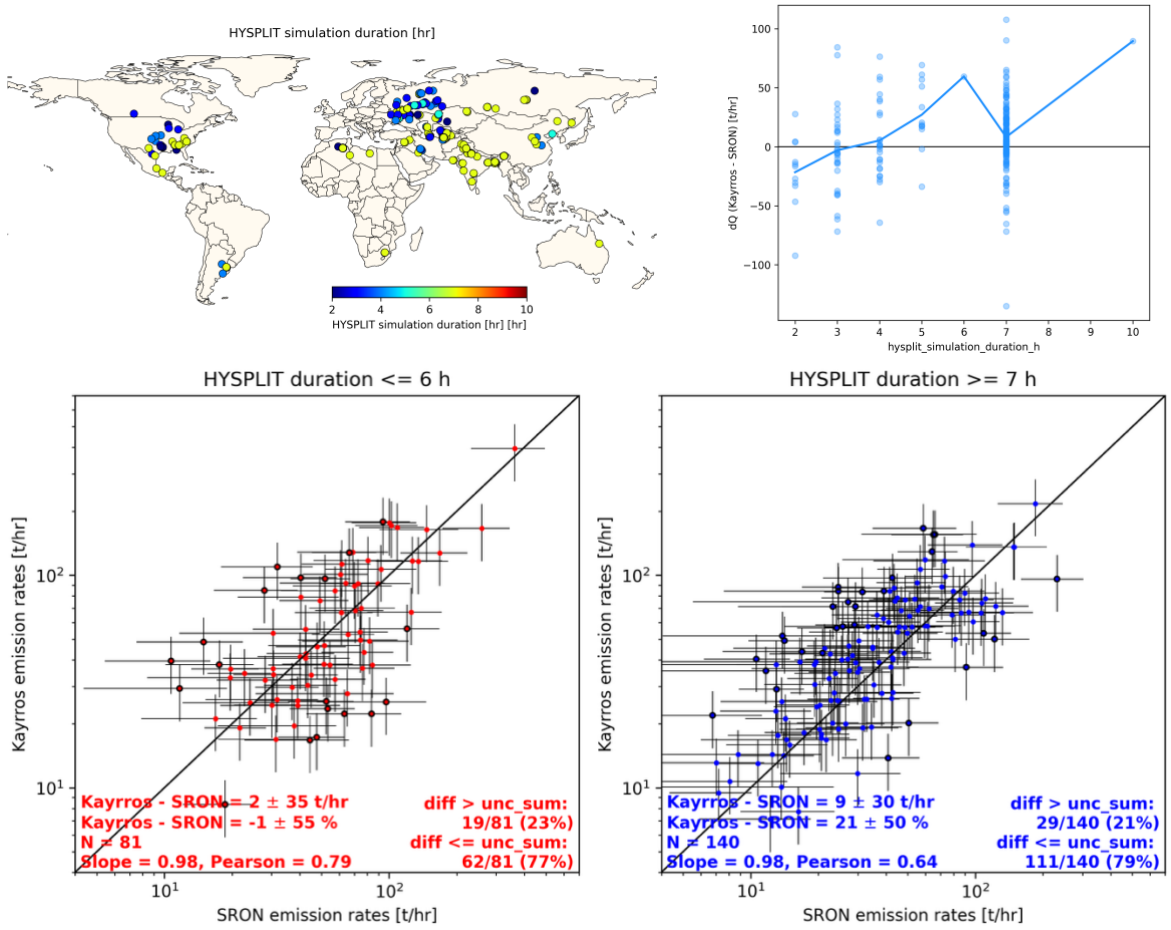


***Figure 5.1.3-4**: Spatial distribution of Kayrros HYSPLIT simulation duration (top left); impact of Kayrros HYSPLIT simulation duration on the Kayrros – SRON emission rate difference (top right); comparison of Kayrros and SRON emission rates for Kayrros HYSPLIT simulation durations below or equal to 6 hours (bottom left) and for durations equal or above 7 hours (bottom right).*

Figure 5.1.3-4 shows the spatial distribution of Kayrros HYSPLIT simulation duration and Kayrros – SRON difference examined for emission durations below or equal to 6 hours, and for emission durations above or equal to 7. Most of the Kayrros-detected plumes quantified with shorter simulated emission durations are located near oil & gas production basins, mainly along pipelines in Russia, or the Permian Basin in the USA. These emission plumes are overall characterized by short duration emission events,

thus explaining why Kayrros HYSPLIT simulations require shorter emission durations to reproduce observations for these cases. **Figure 5.1.3-4**, also shows that shorter events/plumes tend to feature lower and even negative Kayrros – SRON differences, whereas longer ones show the opposite. The scatter of Kayrros – SRON difference is also larger for short events compared to longer ones. This impact of the emission event duration on Kayrros – SRON differences hints at a significant role played by the quantification methods – and how they account for atmospheric transport – to explain these differences.

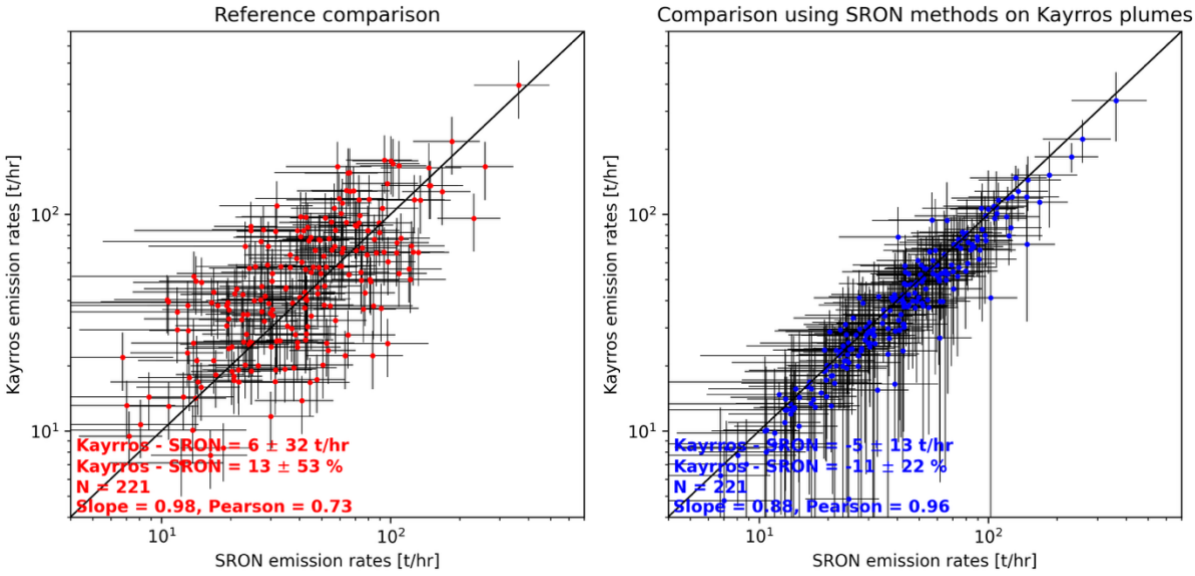


***Figure 5.1.3-5**: Kayrros – SRON emission rate comparison for Kayrros emission rates quantified using the Kayrros quantification method (left) and for Kayrros emission rates quantified by using the SRON quantification method on the Kayrros-detected plumes (plume mask + estimated background XCH4 values). Inset percentages are provided relative to the averaged SRON plume emission rate.*

Reproducing Kayrros emission quantification based on HYSPLIT plume simulations is out of the scope of MEDUSA WP320. In order to discuss the impact of the quantification method differences between Kayrros and SRON, and separate it from the impact of the detected plume masks, we can however apply the SRON quantification methods on Kayrros-detected plume masks and estimated XCH4 background values. **Figure 5.1.3-5** shows the baseline Kayrros – SRON emission comparison where Kayrros-detected plumes are quantified by Kayrros' method and an adjusted comparison where emission rates for Kayrros-detected plumes are quantified using the SRON method. We find that applying the SRON quantification methods on Kayrros detected plumes reduces the Kayrros – SRON difference standard deviation from 32 t/hr to 13 t/hr (variance reduction of 83%). We conclude that the emission rate quantification method (fitting a HYSPLIT simulated plume driven by GFS meteorology, compared to the IME method relying on ERA5 and GEOS-FP) is the main contributor

to the Kayrros – SRON difference scatter. The detected plumes themselves (plume mask and estimated XCH4 background) only cause 17% of the difference variance.

Finally, for the SRON quantification method, we explore the -10% difference between emission rate quantifications obtained based on Kayrros and SRON plumes in **Figure 5.1.3-5**. **Figure 5.1.3-6** compares integrated mass enhancement (in kg), effective plume lengths, and estimated XCH4 background between Kayrros and SRON plumes. We see that Kayrros plumes are 3% shorter than SRON plumes (percentages relative to the averaged plume length over the dataset). This translates into Kayrros plumes having 9% less methane excess mass, which is very close to the -10% emission rate difference obtained for emission rates in **Figure 5.1.3-5** (in the IME method, the emission rate is a linear function of plume mass, which is directly related to area, assuming a plume with constant enhancement, which is the square of the plume length). We can thus conclude that, for one-to-one matching plumes, Kayrros delineates slightly smaller plumes than SRON, resulting in a 10% lower emission rate when the same method is applied.

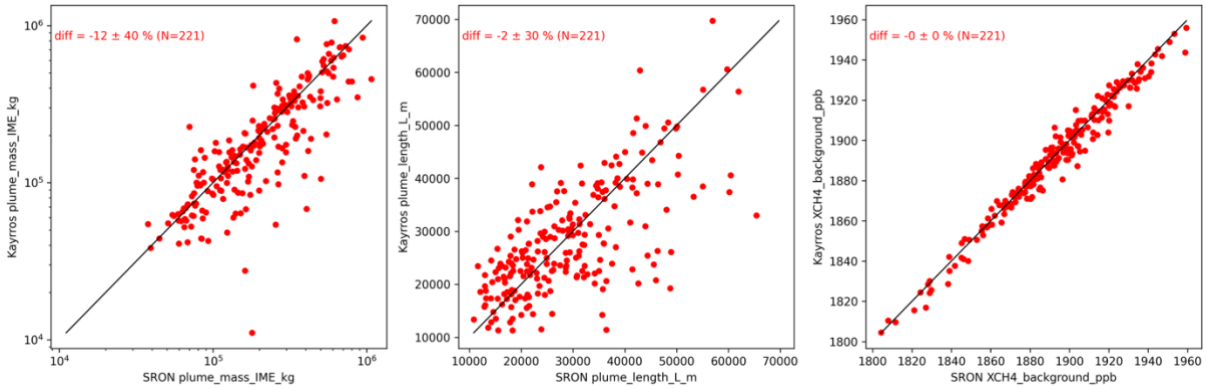


***Figure 5.1.3-6**: Comparison of plume mass (left), plume length (center) and estimated XCH4 background for Kayrros (y-axis) and SRON (x-axis) plumes. Inset differences are given in percentages relative to the average SRON plume mass, plume length, and estimated XCH4 background.*

5.1.4. Results: Plumes detected by Kayrros but not reported by SRON

As provided in **Table 5.1.2-1**, Kayrros detected 386 plumes in 2021 that are not matching any plume reported by SRON. To precise the formulation employed here, we qualify these plumes as “not reported” as to not presume of the reason why these plumes are missing from the other dataset. They could partly be explained by non-detections, but for example also by filtering of detections.

Here, we explore the complete record of TROPOMI data and SRON (reported and discarded) detections to explain these Kayrros detections that have not been reported by SRON. The map in **Figure 5.1.4-1** illustrates the spatial distribution of these plumes missing from the dataset SRON reported.

First, examining TROPOMI data that enter the machine-learning framework used at SRON for plume detection, we find that pre-filtering of TROPOMI pixels before detection is attempted can explain 150/386 (39%) of these missing plumes. These filters are described in Schuit et al. (2023). They are not strict on the quality assurance value (all values above 0.4 are accepted) as to provide more coverage, but are strict on supporting parameters (e.g. albedo, aerosol optical depth, etc.) to avoid retrieval artefacts (e.g. low albedo, snowy surface, high retrieval uncertainty, strong aerosol optical depth, etc.). In comparison, Kayrros allows all pixels with a quality assurance value above 0.75 (Lauvaux et al., 2022).

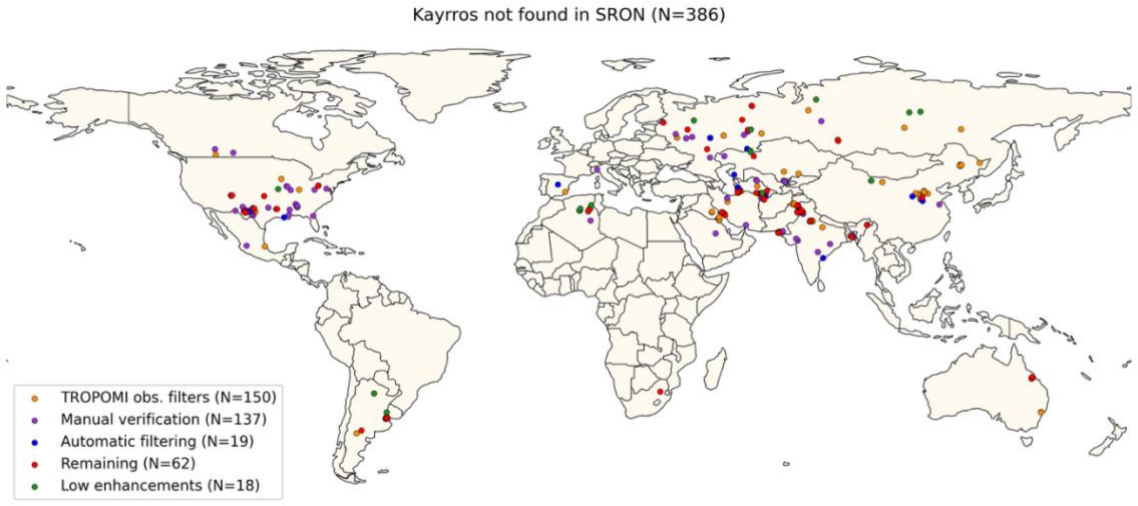


***Figure 5.1.4-1**: Spatial distribution of the 386 plumes detected by Kayrros that were not reported by SRON for 2021, colored by the explanation of why they were not reported by SRON.*

After a detection of plume-like features in TROPOMI data, the SRON machine-learning framework for plume detection includes a second stage of automated artefact filtering (see **Sect. 3.1.1** or in Schuit et al., 2023). We examined the complete SRON-internal record of detections automatically removed by this second stage filtering, and identified 19/386 (5%) plume detections that are reported in the Kayrros dataset but not in SRON's.

After a detection is confirmed by this second stage of automated filtering, it is manually labelled by two human operators, ensuring that no false positives remain in the final dataset. In the end, only clear and confident plumes are reported. Exploring the SRON-internal records of all labelled detections (that passed all automated detection and filtering steps, including both reported and manually-excluded detections), we found that 137/386 (35%) of the plumes included in the Kayrros dataset and not reported by SRON have been detected by SRON, but manually excluded during human labelling, so as to report only clear and confident plumes.

We find that 18/386 (5%) of the Kayrros detections not reported in SRON's dataset show very low enhancements in the latest version of the TROPOMI OFFL data product. As some of the plumes reported by Kayrros can be based on an earlier TROPOMI data version, it is possible that a few plumes reported by Kayrros do not appear anymore in the updated version of the TROPOMI data.

Lastly, 62/386 (16%) plumes detected by Kayrros but not reported by SRON remain to be explained by further studies. These could focus on understanding the behavior of the CNN detecting plume-like features (first part of the SRON detection pipeline, see **Sect 3.1.1**) on these specific cases.

Finally, we also examine in **Figure 5.1.4-2** how Kayrros plume feature distributions (e.g. plume length, plume mass, etc.) computed from the latest TROPOMI data version compare between one-to-one matching plumes and the plumes detected by Kayrros not reported by SRON. The latter set includes a few plumes with lower mass per area than the minimum for one-to-one matching plumes: these most likely correspond to the few plumes discussed above that show very small enhancement, possibly pointing towards a change between data versions. Beyond this small difference, there are no significant shifts between the feature distributions for the one-to-one matching plumes and the ones detected by Kayrros and not reported by SRON.

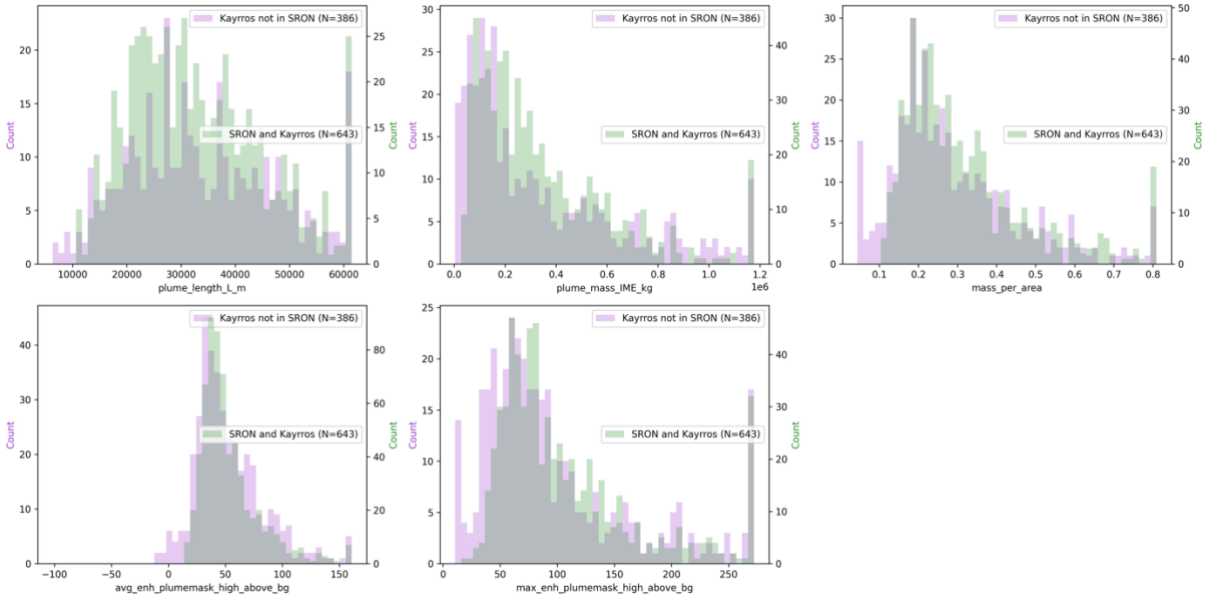


***Figure 5.1.4-2**: Distributions of Kayrros recomputed effective plume length (top left), plume mass (top center), mass per area ratio (top right), plume-averaged methane enhancement (bottom left), plume maximum methane enhancement (bottom center) for the 643 one-to-one Kayrros – SRON plume matches (green) and the 386 plumes detected by Kayrros and not reported by SRON (purple).*

Overall, this analysis of plumes included in Kayrros' dataset and not reported in SRON's has underlined the dominant contribution of filtering in the SRON detection framework (TROPOMI pixel filters before attempting detection, and human labelling: 74% of not reported plumes) to explain the difference. Whether this filtering is indicated pertains to the question of understanding and defining what true/false positives are for each group. A plume-by-plume discussion would need to be organized to explore this question, but this is out of the scope of this study.

5.1.5. Results: Plumes detected by SRON but not reported by Kayrros

As provided in **Table 5.1.2-1**, SRON detected 2568 plumes in 2021 that are not matching any plume reported by Kayrros. The map in **Figure 5.1.5-1** illustrates the spatial distribution of these plumes, that are found across all the locations where methane emission plumes are detected by SRON in TROPOMI data.

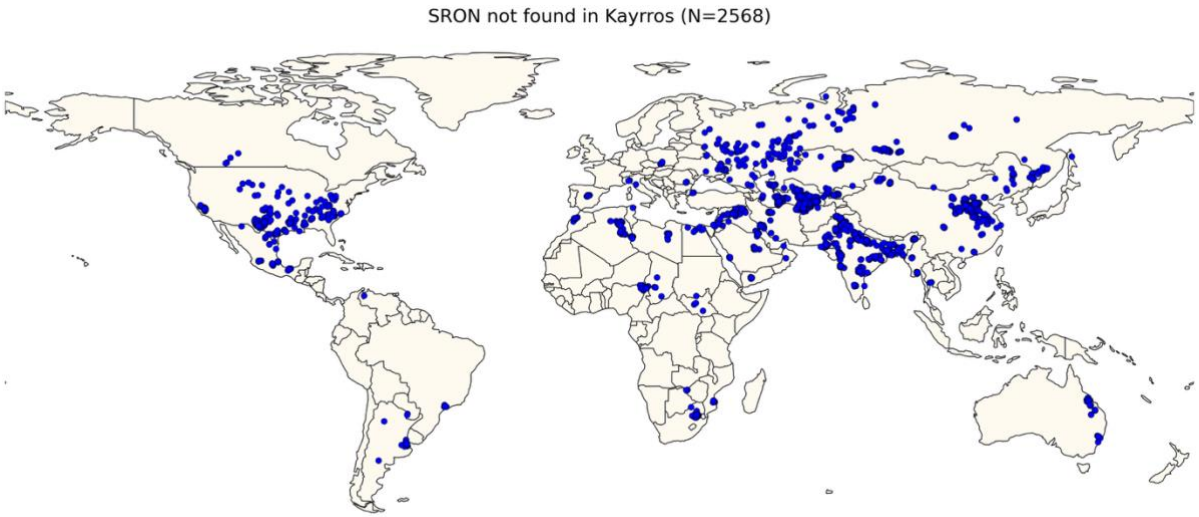


***Figure 5.1.5-1**: Spatial distribution of the 2568 plumes detected by SRON in 2021 TROPOMI data and not reported by Kayrros.*

Among these 2568 plumes, a detailed examination of Kayrros internal records shows that 1787 (70%) plumes were missed by the Kayrros detection model and therefore not assessed further. Of the remaining detections, 398 (15%) plumes were excluded by internal conservative rules applied when manually filtering the output of the model, while 383 (15%) were detected but with non-overlapping plume masks, preventing them from being considered as valid matches. These findings indicate that a large share of SRON-only detections correspond to cases where Kayrros anomaly-based computer vision approach does not identify plumes.

To better understand what characterizes plumes detected by SRON and not reported by Kayrros, **Figure 5.1.5-2** provides the distribution of key plume features computed from SRON plume masks for two different sets: the set of 643 one-to-one Kayrros – SRON matching plumes, and this set of 2568 plumes detected by SRON and not reported by Kayrros. Across all key plume features, the one-to-one Kayrros – SRON matching set of plumes shows distributions on the larger end of these features (e.g. effective plume length, methane mass in the plume, and maximum plume methane enhancement to scene background noise ratio), compared to the SRON-only plumes. Plumes not reported by Kayrros in SRON's dataset are on average shorter, less strongly enhanced and thus 'lighter' than plumes from one-to-one Kayrros – SRON matching set. Finally, **Figure 5.1.5-3** provides a complementary view on these two sets

of plumes considering the cross- 2-dimensional distributions of SRON effective plume length and maximum plume enhancement to scene background noise ratio (two uncorrelated parameters). Overall, plumes detected by SRON and not reported by Kayrros are relatively short and low enhancement plumes. These distributions are most likely explained by the Kayrros plume detection method: it relies on a computer vision framework which requires a preliminary smoothing of TROPOMI data (see methods in Lauvaux et al., 2022) that can dampen the less intense plumes.

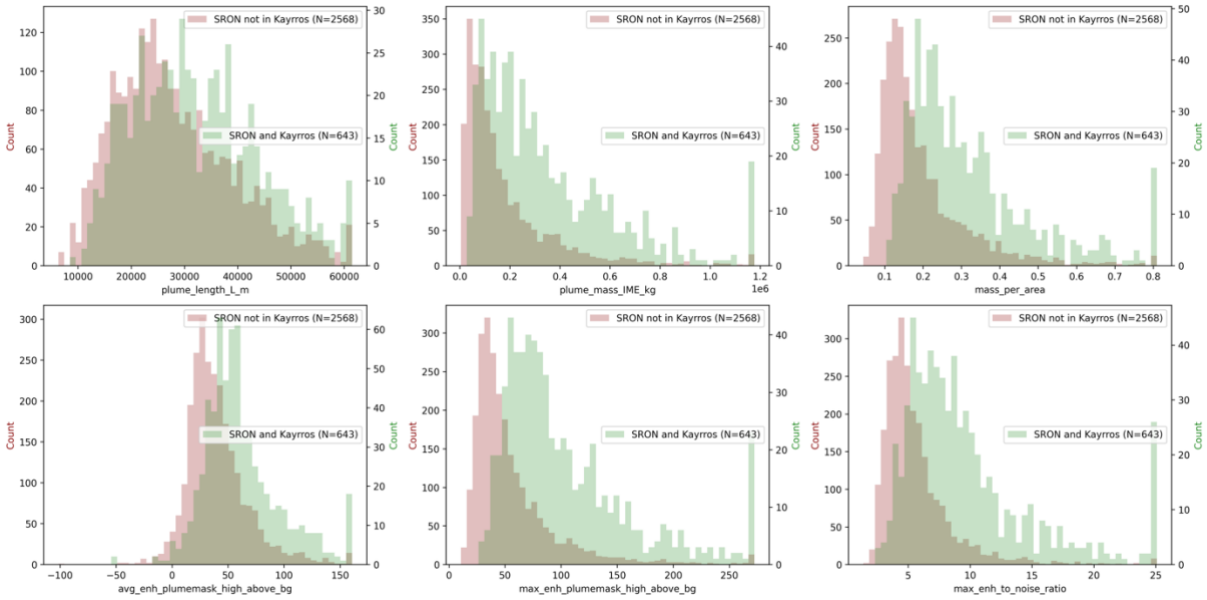


***Figure 5.1.5-2**: Distributions of SRON effective plume length (top left), plume mass (top center), mass per area ratio (top right), plume-averaged methane enhancement (bottom left), plume maximum methane enhancement (bottom center) and maximum plume methane enhancement to scene background noise ratio (bottom right) for the 643 one-to-one Kayrros – SRON plume matches (green) and the 2568 plumes detected by SRON and not reported by Kayrros (red).*

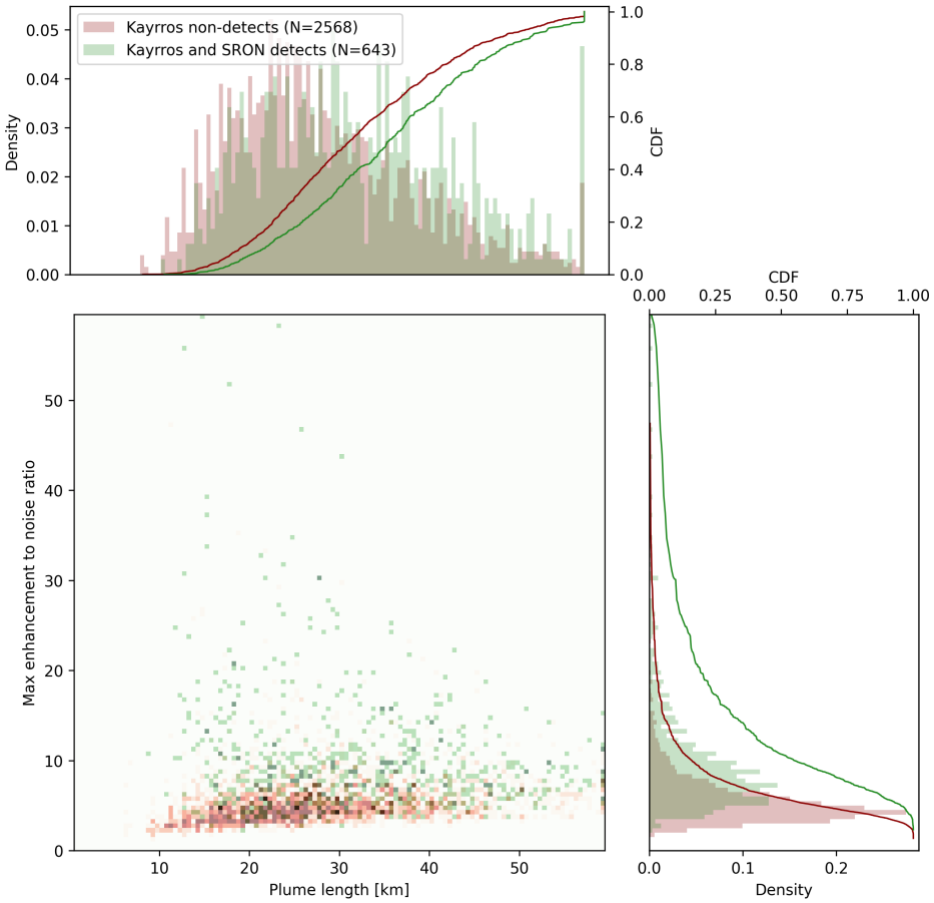


***Figure 5.1.5-3**: One and two-dimensional distributions of one-to-one Kayrros – SRON matching plumes (green) and SRON-only plumes (red) for two uncorrelated key plume features: effective plume length (top, bottom left x-axis) and maximum plume enhancement to scene background noise ratio (bottom right, bottom left y-axis).*

In addition, specific conditions explain why certain SRON detections were discarded by Kayrros when manually filtering to keep only clear and confident plumes. These include scenes with very noisy retrievals, plumes occurring in pixels with marginal albedo values, or plumes located very close to no-data areas in the albedo or Level-2 layers.

Finally, it is worth noting that the differences between the two datasets are also partly linked to the use of different TROPOMI product versions. Some SRON-only plumes fall within no-data regions in the OFFL product used by Kayrros. Reprocessing of Kayrros detections with the RPRO product could therefore provide an opportunity to improve alignment between the two datasets.

5.2. Summary of global comparisons

Global datasets of methane emission plumes detected in TROPOMI data have been reported by Kayrros and SRON for the year 2021. Both datasets have detections covering the main methane emitting regions (e.g Permian basin, West coast of Turkmenistan, Shanxi coal mines, etc.), with SRON also covering additional locations (e.g. Casablanca, Istanbul or coal mines in Poland) consistent with a three-times higher number of detections (3233 SRON versus 1060 Kayrros). Plumes from both datasets have been matched based on the comparison of their reported plume masks. The Kayrros – SRON matching plumes (N=643, with 306 that include emission quantifications in both) agree on average (small mean difference) and show a good correlation (r = 0.62). This result is quite remarkable considering that Kayrros and SRON quantification methods are strictly different (except for the input TROPOMI data). However, the scatter of Kayrros – SRON differences is quite significant, with a 53% relative difference standard deviation. Considering the 30% and 40% mean relative uncertainties reported by Kayrros and SRON (resp.), 77% of the 306 plumes agree within uncertainty range. Most of the plume-to-plume emission difference scatter is shown to be explained by the difference between Kayrros and SRON emission quantification methods: using SRON's quantification method on Kayrros detected plumes reduces difference variance by 83%. This comparison of Kayrros – SRON matching plumes also brought up useful improvement points for both algorithms: SRON's effective wind calibration are slightly inconsistent between 10 m and PBL-averaged wind speeds; Kayrros' handling of missing TROPOMI pixels appears to worsen comparison results. Examining the 386 plumes detected by Kayrros but not reported by SRON, we assess that filtering (39% due to TROPOMI pixel filters and 35% due to human labelling of detections) employed by the SRON TROPOMI plume detection framework can explain the large majority of these non-detected or non-reported plumes. Conversely, of the 2568 SRON-only detections, 70% correspond to short and low-intensity plumes, which Kayrros' anomaly-based computer vision algorithm tends to miss due to its reliance on smoothed input data. Additional differences (15%) stem from borderline cases, where variations in pipeline design and input data can cause detections to fall into different filtering categories across teams.

6. References

Borchardt, J., S. J. Harris, J. M. Hacker, M. Lunt, S. Krautwurst, M. Bai, H. Bösch, H. Bovensmann, J. P. Burrows, S. Chakravarty, R. A. Field, K. Gerilowski, O. Huhs, W. Junkermann, B. F. J. Kelly, M. Kumm, W. Lieff, A. McGrath, A. Murphy, J. Schindewolf, and J. Thoböll, Insights into Elevated Methane Emissions from an Australian Open-Cut Coal Mine Using Two Independent Airborne Techniques, Environmental Science & Technology Letters, DOI: 10.1021/acs.estlett.4c01063, https://doi.org/10.1021/acs.estlett.4c01063, 2025.

Buchwitz, M., Roger, J., Groshenry, A., Schneising-Weigel, O., Hilker, M., Noël, S., Reuter, M., Hachmeister, J., Boesch, H., Dogniaux, M., Zhang, X., Maasakkers, J. D., Aben, I., Gorroño, J., Guanter, L., Algorithm Intercomparison Plan (AIP), WP310 D3.1, version 1.1, Doc ID: SRON-ESG-PL-2024-003, Date: 16-12-2024, pp. 59, 2024.

Buchwitz, M., O. Schneising-Weigel, S. Noël, M. Reuter, M. Hilker, J. Hachmeister, H. Bovensmann, H. Boesch, F. Reuland, A. Brandt, T. Adams and E. A. Kort, Methane emission estimates of localized sources from Sentinel-5 Precursor, PRISMA, EnMAP and EMIT using a cross-sectional-flux method, oral presentation given at 21st International Workshop on Greenhouse Gas Measurements from Space (IWGGMS-21), https://www.nies.go.jp/soc/en/events/iwggms21/, 9-12 June 2025, Takamatsu, Japan, 2025.

Ember DUH, Brief: Urgency to update Germany's coal mine methane emission factor, https://ember-climate.org/insights/in-brief/de-undermines-cmm-emissions/, Ember Climate and Deutsche Umwelthilfe (DUH), including TROPOMI analysis supporting material and press release (https://www.duh.de/fileadmin/user_upload/download/Pressemitteilungen/Energie/Methan/Embargoed_Public_of_Ember_Brief__Urgency_to_update_methane_emission_factors_from_Germany's_coal_mining_sector.pdf), 10-Apr-2024, 2024.

Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., and Conley, S.: Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region, P. Natl. Acad. Sci. USA, 113, 9734–9739, https://doi.org/10.1073/pnas.1605617113, 2016.

Fuentes Andrade, B., Buchwitz, M., Reuter, M., Bovensmann, H., Richter, A., Boesch, H., and Burrows, J. P.: A method for estimating localized CO_2 emissions from co-located satellite XCO_2 and NO_2 images, Atmos. Meas. Tech., 17, 1145–1173, https://doi.org/10.5194/amt-17-1145-2024, 2024.

Gloudemans, A. M. S., Schrijver, H., Hasekamp, O. P., and Aben, I.: Error analysis for CO and CH_4 total column retrievals from SCIAMACHY 2.3 µm spectra, Atmos. Chem. Phys., 8, 3999–4017, https://doi.org/10.5194/acp-8-3999-2008, 2008.

Groshenry, A., Giron, C., Lauvaux, T., d'Aspremont, A. & Ehret, T. Detecting methane plumes using prisma: deep learning model and data augmentation. In

NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning (NIPS, 2022). https://www.climatechange.ai/papers/neurips2022/17, 2022

Groshenry A., Giron C., Hessel C., de Franchis C., Facciolo G., Ehret T., “Methane emissions monitoring using geostationary satellites” in IGARSS 2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024.

Guanter, L., I. Irakulis-Loitxate, J. Gorroño, E. Sánchez-García, D. H. Cusworth, D. J. Varon, S. Cogliati, and R. Colombo, “Mapping methane point emissions with the prisma spaceborne imaging spectrometer,” Remote Sensing of Environment, vol. 265, p. 112671, 2021.

Guanter, Luis, Javier Roger, Shubham Sharma, Adriana Valverde, Itziar Irakulis-Loitxate, Javier Gorroño, Xin Zhang, Berend J. Schuit, Joannes D. Maasakkers, Ilse Aben, Alexis Groshenry, Antoine Benoit, Quentin Peyle and Daniel Zavala-Araiza, Multisatellite Data Depicts a Record-Breaking Methane Leak from a Well Blowout, 30 June 2024, Environmental Science & Technology Letters, DOI: 10.1021/acs.estlett.4c00399, 2024.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, ERA Report Series, https://doi.org/10.21957/tkic6g3wm, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloy- aux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. Roy. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Jacob, D. J., Turner, A. J., Maasakkers, J. D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., and Frankenberg, C.: Satellite observations of atmospheric methane and their value for quantifying methane emissions, Atmos. Chem. Phys., 16, 14371–14396, https://doi.org/10.5194/acp-16-14371-2016, 2016.

Jacob, D. J., Varon, D. J., Cusworth, D. H., et al., Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane, Atmos. Chem. Phys., 22, 9617–9630, doi:10.5194/acp-22-9617-2022, 2022.

Krautwurst, S., Gerilowski, K., Borchardt, J., Wildmann, N., Gałkowski, M., Swolkień, J., Marshall, J., Fiehn, A., Roiger, A., Ruhtz, T., Gerbig, C., Necki, J., Burrows, J. P., Fix, A., and Bovensmann, H.: Quantification of CH_4 coal mining emissions in Upper Silesia by passive airborne remote sensing observations with the Methane Airborne MAPper (MAMAP) instrument during the CO2 and Methane (CoMet) campaign, Atmos. Chem. Phys., 21, 17345–17371, https://doi.org/10.5194/acp-21-17345-2021, 2021.

Krautwurst, S., Fruck, C., Wolff, S., Borchardt, J., Huhs, O., Gerilowski, K., Gałkowski, M., Kiemle, C., Quatrevalet, M., Wirth, M., Mallaun, C., Burrows, J. P., Gerbig, C., Fix, A., Bösch, H., and Bovensmann, H.: Identification and Quantification of CH4 Emissions from Madrid Landfills using Airborne Imaging Spectrometry and Greenhouse Gas Lidar, EGUsphere [preprint] (accepted, in press), https://doi.org/10.5194/egusphere-2024-3182, 2024.

Lauvaux, T.; Giron, C.; Mazzolini, M.; d'Aspremont, A.; Duren, R.; Cusworth, D.; Shindell, D.; Ciais, P. Global Assessment of Oil and Gas Methane Ultra-Emitters. arXiv preprint arXiv:2105.06387 2021.

Lo Piano, Samuele, et al., Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy, Reliability Engineering and System Safety, 206, 107300, https://www.sciencedirect.com/science/article/pii/S0951832020307961 , 2021.

Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., aan de Brugh, J., Schneider, A., Wu, L., Hase, F., Kivi, R., Wunch, D., Pollard, D. F., Shiomi, K., Deutscher, N. M., Velazco, V. A., Roehl, C. M., Wennberg, P. O., Warneke, T., and Landgraf, J.: Methane retrieved from TROPOMI: improvement of the data product and validation of the first 2 years of measurements, Atmos. Meas. Tech., 14, 665–684, https://doi.org/10.5194/amt-14-665-2021, 2021.

Lorente, A., Borsdorff, T., Martinez-Velarte, M. C., and Landgraf, J.: Accounting for surface reflectance spectral features in TROPOMI methane retrievals, Atmos. Meas. Tech., 16, 1597–1608, https://doi.org/10.5194/amt-16-1597-2023, 2023.

Maasakkers, J. D., Varon, D. J., Elfarsdóttir, A., McKeever, J., Jervis, D., Mahapatra, G., Pandey, S., Lorente, A., Borsdorff, T., Foorthuis. L. R., Schuit, B. J.,Tol, P., van Kempen, T. A., van Hees, R., Aben, I., Using satellites to uncover large methane emissions from landfills, Sci. Adv., Vol 8, Issue 32, DOI: 10.1126/sciadv.abn9683, 2022.

Molod, A., Takacs, L., Suarez, M., Bacmeister, J., Song, I.-S., and Eichmann, A.: The GEOS-5 Atmospheric General Circu- lation Model: Mean Climate and Development from MERRA to Fortuna, Technical Report Series on Global Modeling and Data Assimilation, Tech. Rep., https://ntrs.nasa.gov/citations/20120011790 (last access: 25 July 2023), 2012.

Naus, S., J. D. Maasakkers, R. Gautam, M. Omara, R. Stikker, A. K. Veenstra, B. Nathan, I. Irakulis-Loitxate, L. Guanter, S. Pandey, M. Girard, A. Lorente, T.

Borsdorff, and I. Aben, Assessing the Relative Importance of Satellite-Detected Methane Superemitters in Quantifying Total Emissions for Oil and Gas Production Areas in Algeria, Environmental Science & Technology, 57 (48), 19545-19556, DOI: 10.1021/acs.est.3c04746, 2023.

Palmer, P. I., Feng, L., Lunt, M. F., Parker, R. J., Bösch, H., Lan, X., Lorente, A., Borsdorff, T., The Added Value of Satellite Observations of Methane For Understanding the Contemporary Methane Budget. Philos. Trans. R. Soc. A, 379 (2210), 20210106, 2021.

Pei, Zhipeng, Ge Han, Huiqin Mao, Cuihong Chen, Tianqi Shi, Keyi Yang, Xin Ma, Wei Gong, Improving quantification of methane point source emissions from imaging spectroscopy, Remote Sensing of Environment, Volume 295, 113652, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2023.113652, 2023.

Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., Bovensmann, H., and Burrows, J. P.: Towards monitoring localized CO2 emissions from space: co-located regional CO_2 and NO_2 enhancements observed by the OCO-2 and S5P satellites, Atmos. Chem. Phys., https://www.atmos-chem-phys.net/19/9371/2019/, 19, 9371-9383, 2019.

Roger, J., Irakulis-Loitxate, I., Valverde, A., Gorrono, J., Chabrillat, S., Brell, M., Guanter, L.,: High-Resolution Methane Mapping With the EnMAP Satellite Imaging Spectroscopy Mission. - IEEE Transactions on Geoscience and Remote Sensing, 62, 1-12, https://doi.org/10.1109/TGRS.2024.3352403, 2024.

Sadavarte, P., Pandey, S., Maasakkers, J. D., Lorente, A., Borsdorff, T., Denier van der Gon, H., Houweling, S., and Aben, I.: Methane Emissions from Superemitting Coal Mines in Australia Quantified Using TROPOMI Satellite Observations, Environ. Sci. Technol., 55, 16573–16580, https://doi.org/10.1021/acs.est.1c03976, 2021.

Sherwin, E. D., Rutherford, J. S., Chen, Y., Aminfard, S., Kort, E. A., Jackson, R. B., and Brandt, A. R.: Single-blind validation of space-based point-source methane emissions detection and quantification, Scientific Reports, 13, 3836, https://doi.org/10.1038/s41598-023-30761-2, 2023.

Sherwin, E. D., El Abbadi, S. H., Burdeau, P. M., Zhang, Z., Chen, Z., Rutherford, J. S., Chen, Y., and Brandt, A. R.: Single-blind test of nine methane-sensing satellite systems from three continents, Atmos. Meas. Tech., 17, 765–782, https://doi.org/10.5194/amt-17-765-2024, 2024.

Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., and Burrows, J. P.: Severe Californian wildfires in November 2018 observed from space: the carbon monoxide perspective, Atmos. Chem. Phys., 20, 3317-3332, https://doi.org/10.5194/acp-20-3317-2020, 2020a.

Schneising, O., Buchwitz, M., Reuter, M., Vanselow, S., Bovensmann, H., and Burrows, J. P.: Remote sensing of methane leakage from natural gas and petroleum systems revisited, Atmos. Chem. Phys., 20, 9169-9182, https://doi.org/10.5194/acp-20-9169-2020, 2020b.

Schneising, O., Buchwitz, M., Hachmeister, J., Vanselow, S., Reuter, M., Buschmann, M., Bovensmann, H., and Burrows, J. P.: Advances in retrieving XCH_4 and XCO from Sentinel-5 Precursor: improvements in the scientific TROPOMI/WFMD algorithm, Atmos. Meas. Tech., 16, 669–694, https://doi.org/10.5194/amt-16-669-2023, 2023.

Schneising, O., Buchwitz, M., Reuter, M., Weimer, M., Bovensmann, H., Burrows, J. P., and Bösch, H.: Towards a sector-specific CO/CO_2 emission ratio: Satellite-based observation of CO release from steel production in Germany, Atmos. Chem. Phys., 24, 7609–7621, https://doi.org/10.5194/acp-24-7609-2024, 2024.

Schuit, B. J., Maasakkers, J. D., Bijl, P., Mahapatra, G., van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D. J., McKeever, J., Jervis, D., Girard, M., Irakulis-Loitxate, I., Gorroño, J., Guanter, L., Cusworth, D. H., and Aben, I.: Automated detection and monitoring of methane super-emitters using satellite data, Atmos. Chem. Phys., 23, 9071–9098, https://doi.org/10.5194/acp-23-9071-2023, 2023.

Stein, A.F., Draxler, R.R, Rolph, G.D., Stunder, B.J.B., Cohen, M.D., and Ngan, F., (2015). NOAA's HYSPLIT atmospheric transport and dispersion modeling system, Bull. Amer. Meteor. Soc., 96, 2059-2077, http://dx.doi.org/10.1175/BAMS-D-14-00110.1, 2015.

Thompson, D. R., Thorpe, A. K., Frankenberg, C., Green, R. O., Duren, R., Guanter, L., Hollstein, A., Middleton, E., Ong, L., and Ungar, S.: Space-based remote imaging spectroscopy of the Aliso Canyon CH4 superemitter, Geophys. Res. Lett., 43, 6571–6578, https://doi.org/10.1002/2016GL069079, 2016.

Thorpe, A. K., Frankenberg, C., and Roberts, D. A.: Retrieval techniques for airborne imaging of methane concentrations using high spatial and moderate spectral resolution: application to AVIRIS, Atmos. Meas. Tech., 7, 491–506, https://doi.org/10.5194/amt-7-491-2014, 2014.

Tu, Q., Hase, F., Schneider, M., García, O., Blumenstock, T., Borsdorff, T., Frey, M., Khosrawi, F., Lorente, A., Alberti, C., Bustos, J. J., Butz, A., Carreño, V., Cuevas, E., Curcoll, R., Diekmann, C. J., Dubravica, D., Ertl, B., Estruch, C., León-Luis, S. F., Marrero, C., Morgui, J.-A., Ramos, R., Scharun, C., Schneider, C., Sepúlveda, E., Toledano, C., and Torres, C.: Quantification of CH_4 emissions from waste disposal sites near the city of Madrid using ground- and space-based observations of COCCON, TROPOMI and IASI, Atmos. Chem. Phys., 22, 295–317, https://doi.org/10.5194/acp-22-295-2022, 2022.

Tu, Q., Hase, F., Qin, K., Cohen, J. B., Khosrawi, F., Zou, X., Schneider, M., and Lu, F.: Quantifying CH_4 emissions from coal mine aggregation areas in Shanxi, China, using TROPOMI observations and the wind-assigned anomaly method, Atmos. Chem. Phys., 24, 4875–4894, https://doi.org/10.5194/acp-24-4875-2024, 2024.

Vanselow, S., Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., Boesch, H., and Burrows, J. P.: Automated detection of regions with persistently enhanced

methane concentrations using Sentinel-5 Precursor satellite data, Atmos. Chem. Phys., 24, 10441–10473, https://doi.org/10.5194/acp-24-10441-2024, 2024.

Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., et al., Satellite discovery of anomalously large methane point sources from oil/gas production. Geophysical Research Letters, 46, 13,507–13,516. https://doi.org/10.1029/2019GL083798, 2019.

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, Atmos. Meas. Tech., 11, 5673–5686, https://doi.org/10.5194/amt-11-5673-2018, 2018.

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 Sens. Environ., 120, 70–83, https://doi.org/10.1016/j.rse.2011.09.027, 2012.

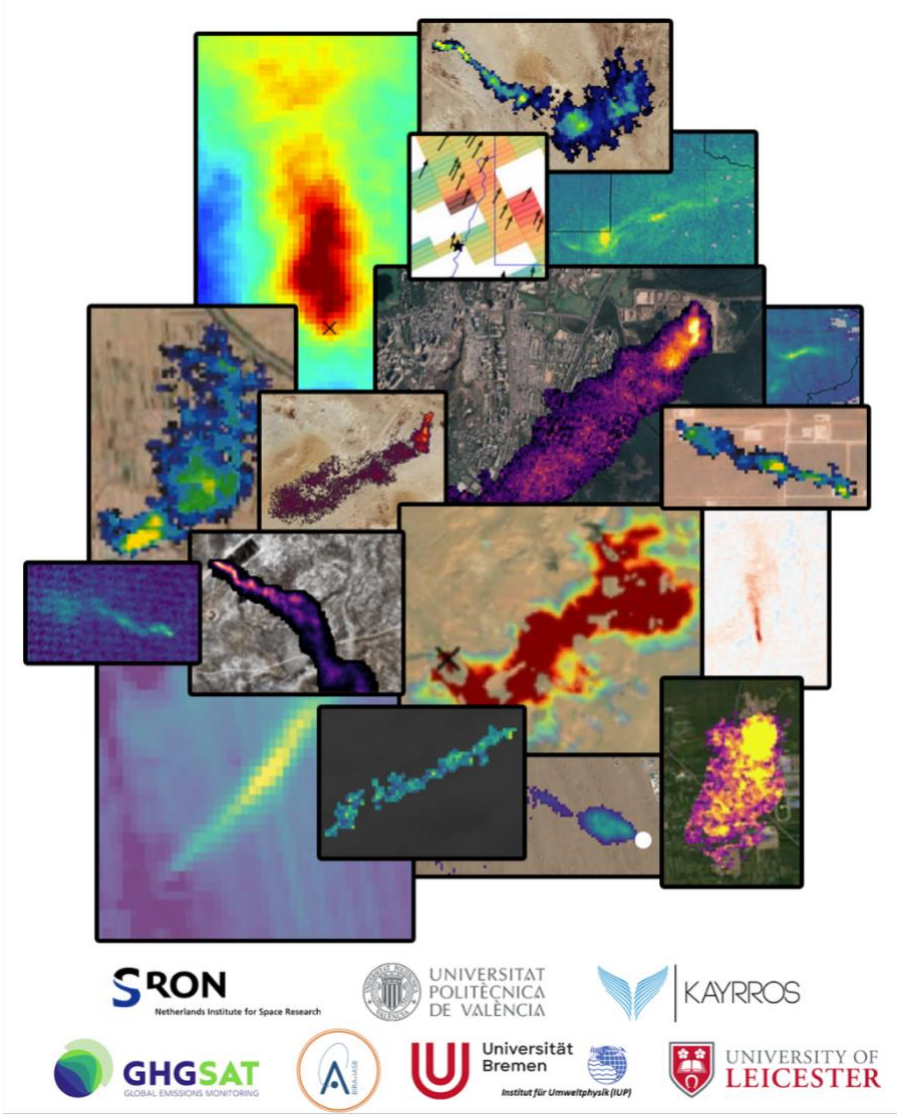
Worden, J., Green, P., Eldering, A. and Sherwin, E., Common Practices for Quantifying Methane Emissions from Plumes Detected by Remote Sensing, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.IR.8575, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=960016 (Accessed July 3, 2025), 2025.

Zhang, Y., Gautam, R., Pandey, S., Omara, M., Maasakkers, J. D., Sadavarte, P., Lyon, D., Nesser, H., Sulprizio, M. P., Varon, D. J., Zhang, R., Houweling, S., Zavala-Araiza, D., Alvarez, R. A., Lorente, A., Hamburg, S. P., Aben, I., and Jacob, D. J.: Quantifying methane emissions from the largest oil-producing basin in the United States from space, Science Advances, 6, eaaz5120, https://doi.org/10.1126/sciadv.aaz5120, 2020.

***** End of Document *****

ANNEX B (HI comparisons) to Algorithm Intercomparison Report (AIR)

WP320 – Annex for Deliverable D3.2



Authors

Institute of Environmental Physics, University of Bremen, Germany (IUP-UB)

- Michael Buchwitz
- Oliver Schneising-Weigel
- Michael Hilker
- Stefan Noël
- Maximilian Reuter
- Jonas Hachmeister
- Hartmut Boesch

SRON Netherlands Institute for Space Research

- Matthieu Dogniaux
- Xin Zhang
- Joannes D. Maasakkers
- Ilse Aben

Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, Spain (UPV)

- Javier Roger
- Luis Guanter

Kayrros, Paris, France

- Maxence Nevoret
- Julian Akani Guery

Table of Contents

1. Executive summary

This document is Annex B from the Algorithm Intercomparison Report (AIR) of ESA project MEDUSA. Annex B focuses on the algorithm intercomparisons among participating groups, specifically on those algorithms applied to the data from the Hyperspectral Imagers (HI), i.e. EnMAP, PRISMA, and EMIT. Intercomparisons are mainly carried out following the Algorithm Intercomparison Plan (AIP).

This document provides in addition a short overview of the satellite sensors used and the methane emission inversion algorithms. The term “algorithm intercomparison” as used in this document primarily refers to comparisons of outputs as generated by different (emission) algorithms as applied to different (satellite) input data sets. These outputs are primarily the derived emission estimates and their uncertainty, but comparisons also extend to additional data such as the underlying atmospheric methane fields, wind speed, etc. Comparison of some of these outputs will be carried out following the AIP to understand differences in the groups’ algorithm pipelines. We can split the intercomparisons in 4 different parts:

- **Part 1**: plume identification and attribution to source (section 4.1)
- **Part 2**: assessment of methane concentration retrievals used for detection (section 4.2)
- **Part 3**: general methane emission quantification assessment (section 4.3)
- **Part 4**: detailed methane emission quantification assessment (section 4.4). Note that algorithms used for this part could have changed in reference to those used in Parts 1-3 due to group algorithm updates.

2. Hyperspectral imagers (HI)

This section provides a short overview of the satellite sensors relevant for the MEDUSA HI algorithm intercomparison. These are sensors where more than one methane emission inversion algorithm is used and evaluated within MEDUSA.

2.1. PRISMA

PRISMA (PRecursore IperSpettrale della Missione Applicativa) is a hyperspectral imaging mission operated by the Italian Space Agency (ASI). Launched in March 2019, PRISMA carries a push-broom hyperspectral sensor capable of capturing images in 239 spectral bands from 400 nm to 2500 nm. The instrument provides detailed spectral information with a spectral resolution and sampling better than 12 nm. The spatial resolution is 30 m for the hyperspectral bands and 5 m for the panchromatic band. PRISMA is mainly operated by user requests for data acquisition. The high spectral resolution across a broad wavelength range allows for detailed analysis of material composition, while the panchromatic band enhances spatial details and aids in sharpening hyperspectral imagery.

2.2. EnMAP

EnMAP (Environmental Mapping and Analysis Program) is a German hyperspectral imaging mission managed by the German Aerospace Center (DLR). Launched in April 2022, EnMAP features a hyperspectral imager with a push-broom sensor that captures data in 242 spectral bands across the VNIR (420–1000 nm) and SWIR (900–2450 nm) ranges. The spectral resolution and sampling are approximately 6.5 nm in the VNIR range and around 10 nm in the SWIR range. Users can request EnMAP data acquisitions for specific areas of interest by data tasking. The high signal-to-noise ratio and continuous coverage from VNIR to SWIR enable precise detection of surface features and materials, facilitating comprehensive environmental monitoring and analysis.

2.3. EMIT

EMIT is an imaging spectrometer mission developed by NASA's Jet Propulsion Laboratory (JPL). Launched in July 2022, EMIT initially focused on mapping the composition of mineral dust in arid regions to better understand its effects on climate. The instrument captures data in the VNIR to SWIR range (380–2500 nm) with hundreds of narrow spectral bands. The spectral resolution and sampling are approximately 7.5 nm, while the spatial resolution is 60 m. EMIT's revisit time depends on its position on the International Space Station (ISS) and the specific target areas but generally ranges from a few days to weeks.

3. Algorithm description

3.1. SRON

Algorithm for Part 1-3
We use L1 data for all instruments. For PRISMA, these are obtained from PRISMA Portal (https://prisma.asi.it/); For EnMAP the L1 data come from EOWEB GeoPortal (https://eoweb.dlr.de/); and for EMIT we download L1 data from NASA Earthdata (https://search.earthdata.nasa.gov/search?q=C2408009906-LPCLOUD). No pre-treatment is applied to the L1 data.

The SRON algorithm applies matched filters with two different wavelength windows, 2110–2450 nm and 1300–2500 nm, to retrieve the methane enhancement and denoised methane enhancement field, respectively. As proposed by Roger et al. (2024), the original methane enhancement field is used for quantification, while the denoised one is used for plume masking. The matched filter is applied per-column, with a land/ocean mask applied based on GSHHS data.

To determine the unit methane absorption spectrum (k), we employ a forward model (Gloudemans et al. 2008) and convolve the transmittance with the imager's central wavelength and FWHM. The k value (ppb^{-1}) for each band is calculated as the slope of the regression between the natural logarithm of transmittance and methane enhancements.

Plume detection is carried out by visual inspection. Then, a Chambolle total variance denoising filter (Chambolle et al. 2004) is applied to obtain a smoothed methane enhancement field from the denoised methane field (generated using the 1300 nm – 2500 nm window). Plume pixels are identified by a 98th percentile mask over land and then smoothed by the median and Gaussian filters. Finally, the retrieval is run again with a lognormal matched filter (Pei et al. 2023) while removing plume pixels from the background estimation to make sure the sparsity assumption of the matched filter is not broken.

The algorithm uses the Integrated Mass Enhancement (IME) method to quantify the emission rate. Following Maasakkers et al. (2022), the effective wind speed is defined as $0.34U_{10} + 0.66$ and $0.34U_{10} + 0.42$ for area sources (landfill) and point sources (Oil & Gas, coal mining), respectively. A more instrument-specific quantification calibration is in development. The primary choice for wind is the ERA5 10-m wind speed. However, the GEOS Forward Processing (GEOS-FP) data is used in cases where the ERA5 wind direction differs from the plume direction. If both the ERA5 and GEOS-FP wind data fail to accurately capture the wind direction, the algorithm defaults to using the ERA5 wind data. The length of the plume is calculated as the square root of the plume's area.

We consider sources of uncertainty in our emission estimations: wind speed error, random error of the retrieval, and uncertainty in the IME calibration. The wind error is set as 1.5 m s^{-1} for wind speeds higher than 3 m/s. For wind speeds lower than 3 m/s, a relative wind error of 50% is applied (Varon et al. 2018). To quantify the effects of random retrieval error, the plume mask is applied to the non-plume pixels across the entire scene and the standard deviation of the emission rates is calculated for those non-plume pixels. To capture the uncertainty in the calibration, the difference between using the point source and area source calibrations is used. Finally, the total uncertainty is defined as the square root of the sum of the squares of the individual uncertainties.

The EnMAP L2 product is orthorectified to mitigate image distortion caused by variable topography by combining Rational Polynomial Coefficients (RPCs) and a Digital Elevation Model (DEM), while the EMIT L2 product is orthorectified using a geometric lookup table (GLT). In contrast, the PRISMA L2 product is not orthorectified due to the absence of RPCs and GLT data in its L1 data. We provide our output in NetCDF format for spatial variables like the plume mask and in CSV files for the source rate and its uncertainty (given in ton/hr).

Algorithm updates for Part 4
Several updates were made for the algorithm version used in Part 4 (section 5.4) of the comparisons.

1) Land/water mask
    a. Old method: GSHHS data.
    b. New method: Integrated data from both OpenStreetMap (OSM) and ESA WorldCover databases (https://hyp3-docs.asf.alaska.edu/water_masking/) for improved accuracy.
2) Denoising
    a. Old method: We applied a Chambolle total variance (TV) denoising filter with constant weights (PRISMA: 150; EMIT/EnMAP: 50) to the methane field generated from the 1300 nm – 2500 nm spectral window.
    b. New method: We apply a TV filter calibrated by J-Invariance to the same methane field.
3) Plume masking
    a. Old method: We generated a 98th percentile mask over land, followed by median and Gaussian smoothing.
    b. New method: We apply a watershedding technique (Zhang et al, 2025) to the denoised methane field. Thresholds of 2 and 3 standard deviations are used to identify multiple localized high-enhancement features and nearby areas with high enhancement values. We dilate these masks by 180 m and merge overlapping masks, with the mask containing the emission source used to identify masks from a single source. To ensure that detected plumes originate from the same source, we constrain the azimuth difference of the oriented envelope (minimum rotated rectangle) to less than 30°, assuming minimal wind direction changes near the emitting source. Non-detects are classified when no plume mask is detected near the source of interest.
4) Lognormal matched filter
    a. Old method: Applied to all cases.
    b. New method: Only applied to emission rates higher than 10 t/h.
5) Wind calibration
    a. Old method: Following Maasakkers et al., 2022, the effective wind speed is defined as $0.34U_{10} + 0.66$ and $0.34U_{10} + 0.42$ for area sources (landfill) and point sources (Oil & Gas, coal mining), respectively.
    b. New method: New instrument-specific wind calibrations based on the updated plume masking method and instrument-specific backgrounds.
        i. Point-source:
            1. EMIT: $0.35U_{10} + 0.43$
            2. EnMAP: $0.32U_{10} + 0.45$
            3. PRISMA: $0.31U_{10} + 0.43$
        ii. Area-source
            1. EMIT: $0.39U_{10} + 0.63$
            2. EnMAP: $0.37U_{10} + 0.68$
            3. PRISMA: $0.37U_{10} + 0.69$
6) Uncertainty of wind calibration
    a. Old method: To capture the uncertainty in calibration, the difference between using the point source and area source calibrations is used.

b. New method: We apply the point-source calibration instead and examine the resulting variations in emission rates. For point sources, the average fit residual is incorporated into the analysis.

Another method to calculate the emission rate is the cross-sectional flux (CSF) method. We provide CSF results to show how different parts of long plumes give different emission rates. The source rate is estimated as the product of the cross-plume gas enhancement integral and a different effective wind speed ($U_{eff,CSF}$):

$$Q = U_{eff,CSF} \int_a^b \Delta\Omega(x,y)dy$$

Here, the x-axis aligns with the wind direction, while the y-axis is oriented perpendicular to it. The integral is evaluated between the plume boundaries [a, b], as defined by the cross section in the plume mask. A centerline curve is fitted to the plume mask using the ddeq library (Kuhlmann, 2024), which constructs a two-dimensional curve based on the pixels within the plume area (Kuhlmann, 2020). The space between each line is defined as 1.5 times the pixel resolution. The column mass enhancement $\Delta\Omega$ (kg m^{-2}) is derived from the methane enhancement (ΔCH_4):

$$\Delta\Omega = \frac{M_{CH_4}}{M_a}\Omega_a \Delta CH_4$$

where M_{CH4} and M_a are the molar masses (kg mol^{-1}) of methane and dry air, and Ω_a is the column of dry air (kg m^{-2}).

Wind data and interpolation

The default wind product used in the analysis is the European Centre for Medium-Range Weather Forecasts Reanalysis 5 (ERA5, 0.25° × 0.25°) 10-m hourly wind data. The GEOS Forward Processing (GEOS-FP, 0.25° latitude × 0.3125° longitude) 10-m wind data is also supported by our methodology. We use the GEOS-FP data in cases where the ERA5 wind direction differs from the plume direction by more than 90 degrees. If both the ERA5 and GEOS-FP wind data fail to accurately capture the wind direction, we default to using the ERA5 wind data.

We use linear interpolation (along the time, latitude, and longitude dimensions) to obtain 2D wind data across the entire scene and compute the mean wind speed within each plume mask.

3.2. Kayrros

Kayrros algorithm is based on the matched filter formulation described in Guanter et al., 2021, applied to the L1B hyperspectral product. The matched filter is performed on a per-column basis on the spectral ranges 1550 nm - 1800 nm and 2100 nm – 2480 nm where methane respectively has weak and strong absorption features.

For every scene, a high-resolution reference spectrum is generated by integrating the Beer-Lambert law on an atmosphere model of homogeneous layers using:

- The observation parameters (viewing zenith angle, solar zenith angle) of the scene.
- The absorption coefficients from the HITRAN 2020 molecular spectroscopic database (Gordon et al., 2022).
- The 1976 US standard atmosphere from LOWTRAN 7 atmospheric transmittance and background radiance model (Kneizys et al., 1988) is used for defining the composition of the atmospheric layers.

The unit absorption spectrum is eventually computed as the Jacobian of the reference absorption spectrum for an additional ppm of methane.

For EMIT, plume detection and delineation are automatically performed by a Deep Learning model. The detections are systematically cross validated by two human expert labelers to filter out any false positive detection by checking the presence of a potentially emitting source, and the consistency of the detection in terms geometric features, wind alignment, albedo or presence of spectral confusers. For PRISMA and EnMAP, plume detection and delineation are performed by two human expert labelers.
For all sensors, the binary masks produced by their respective detection methods are refined by morphological operations and intersection with the result of a t-test with the background noise. If the RPCs are provided in the L1B product, the detections are coregistered by iteratively refining the projection ellipsoid with an appropriate DEM.

Every detection is quantified using the Integrated Mass Enhancement (IME) method (Varon et al., 2018). The ERA5 10 m wind speed reanalysis data (Hersbach et al., 2018) are used, assuming a default 0.5 m s^{-1} uncertainty for each component of the wind vector. The closest grid point to the plume source is selected, using the last available time step before the acquisition start time. No temporal or spatial interpolation is applied. The wind speed is converted into effective wind speed according to the formula from Guanter et al., 2021.

*Ueff = 0.34 * U10 + 0.44.*

A flux rate uncertainty is estimated by combining the uncertainty on the wind data and an empirical 1-σ estimate of the retrieval noise.

3.3. UPV

We use radiance L1 data for the PRISMA, EnMAP, and EMIT missions. Data from PRISMA and EnMAP can be acquired by tasking or from the archive, using the PRISMA (https://prisma.asi.it/) and EnMAP portals (https://planning.enmap.org/ips/app), respectively. On the other hand, EMIT data is acquired from the EMIT L1B product from the NASA's EarthData portal (https://search.earthdata.nasa.gov/search). Note that we do not perform any pretreatment of the radiance data.

We filter out the radiance values related to clouds and their shadows. We do the same with water if potential plumes are in land, and with the land pixels if they are in water. We identify these pixels by a manually selected radiance threshold set by visual inspection. We also filter out negative values, as they do not contain any physical meaning.

The methodology used to retrieve methane concentration enhancement maps is the matched-filter (MF) method and related flavors. For plume detection, we use the Combo-MF as in Roger et al., 2024, although we use the simple matched-filter as in Guanter et al., 2021, for emissions coming from area sources, since the emission enhancement is generally more penalized by the Combo-MF due to the underlying surface. On the other hand, for quantification, we use an in-house lognormal MF (LMF) based on Pei et al., 2023, which considers two LMF retrievals that use different unit methane absorption spectra (K), one obtained considering a range of methane concentration enhancement values (0 – 1000 ppb), and the other considering a wider range (0 – 2000 ppb). The first one is more accurate at lower values, while the other is more accurate at higher values. Retrieval values from the first retrieval greater than 1000 ppb are substituted by the second retrieval values.

K is deduced using a lookup table that relates methane transmittance spectra obtained using the MODTRAN radiative transfer code to different values of methane column mixing ratio, while accounting for the specific angular configuration at measurement time. For this task, we configure MODTRAN in spectral transmittance mode assuming atmospheric profiles from the U.S. Standard Atmosphere. Then, K is obtained as the slope from a linear fit applied to relate methane concentration enhancement values to changes in the convolved-to-the-instrument transmittance spectra. Moreover, the calculation of K for satellite-based instruments also assumes an integration over an 8-km high column (Thompson et al., 2016).

Regarding the spectral window of application of the applied methodology, we use 2100 – 2400 nm as in Guanter et al., 2021, for every mentioned method, except for the Combo-MF, which uses a 1000 – 2500 nm spectral range. Note that in the latter window spectral bands with very high-water vapor absorption are removed because sometimes they contain pixels with zero value (see Roger et al., 2024).

The retrieval methods are applied in a per-column basis to avoid effects from the non-uniformity from the across-track detectors from the instruments. In addition, statistics for our retrieval methods are obtained filtering out those pixel values greater than 2-standard deviations in a first run to respect the sparsity assumption from the matched filter.

Plume detection is made by visual inspection in those retrievals used for detection. The potential plume must 1) follow wind direction, 2) originate from a potential emitting source, and 3) have a plume-like shape. Wind speed values are obtained from GEOS-FP reanalysis product (Molod et al., 2012), from which we obtain U10, the wind speed at 10 m above the

surface. First, we search the values of the U10 vertical and horizontal components related to the GEOS-FP pixel that is closer to the latitude and longitude coordinates from the instrument acquisition central coordinates. For each component, we extract the values immediately before and after the acquisition time, which have a 1 hour temporal distance due to the hourly instrument temporal resolution. Next, we linearly interpolate these values to 10 minutes before the acquisition time. The wind speed at this time will have more impact in the plume dynamics rather than the one at the time of the acquisition. Finally, we obtain the U10 by calculating the square root of the quadratic sum of the interpolated components. On the other hand, plume delineation is also based on those retrievals used for detection. The plume mask is obtained manually by combining the visual screening of the retrieval with the definition of the plume polygon through mouse clicks.

The quantification is made using the Integrated Mass Enhancement (IME) method, from which we obtain the flux rate (t/h) emission estimate. We select an effective wind speed calibration of $0.34 \cdot U10 + 0.44$ for point-source emissions captured by the PRISMA and EnMAP instruments (Guanter et al., 2021), and $0.3 \cdot U10 + 0.47$ for point-source emissions captured by the EMIT instrument (Guanter et al., 2024). On the other hand, we select a calibration of $0.34 \cdot U10 + 0.42$ for area-source emission for the three missions (Maasakkers et al., 2022). In addition, the plume length (L) is obtained from the square root of the masked plume area (Varon et al., 2018).

The flux rate emission uncertainty is obtained by the square root of the quadratic sum of the errors related to the IME and the Ueff values. The latter error is driven mainly by the error related to U10. Specifically, we consider an U10 error of 2 m/s if $U10 > 4$ m/s, and an error of $0.5 \cdot U10$ if $U10 < 4$ m/s (Carvalho et al., 2019).

For the PRISMA case, methane retrievals are co-registered by a reprojection applied following a number of GCPs extracted from the L1 metadata. However, the EnMAP and EMIT methane retrievals are orthorectified. A combination of Rational Polynomial Coefficients (RPC) from the metadata and a Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM) instrument are used for the EnMAP methane retrievals. On the other hand, the Geometric Lookup Tables (GLT) from the metadata are used to orthorectify the EMIT methane retrievals.

This algorithm was used by UPV for every intercomparison activity within this work.

3.4. IUP-UB

The algorithm used by IUP-UB to obtain methane emission estimates from the Hyperspectral Imagers (HI) PRISMA, EnMAP and EMIT consists of two main processing steps:

- Level 1 (L1) to Level 2 (L2) processing or "retrieval": Converts L1 input data, i.e., the geolocated radiance spectra, into L2 data products, i.e., to images of atmospheric methane enhancements, ΔXCH4, where XCH4 is the column-averaged dry-air mole fraction of methane (in parts per billion, ppb, or parts per million, ppm).
- Level 2 to Level 4 (L4) processing or "inversion": Converts the L2 product into a L4, i.e., methane emission estimation data product.

These 2 main processing steps are shortly described in the following.

L1 to L2 processing (retrieval):

Input data for the retrieval algorithm are the Level 1B (L1B) satellite data products from PRISMA, EnMAP and EMIT. For the results presented here we use the Matched Filter (MF) method (Bösch and Hilker, 2024) to retrieve ΔXCH4 from data contained in the L1B files. The Matched Filter works on the logarithm of the radiance in a fit window that contains the 2.3μm CH4 absorption band.

As shown in Bösch and Hilker, 2024, several retrieval methods are currently under investigation at IUP-UB, but preliminary analysis suggest that the current implementation of the MF method gives often the best results. The IUP-UB MF retrievals are therefore used as input for the emission estimation step described in the following.

L2 to L4 processing (inversion):

Input data for this algorithm step are the methane enhancement images as obtained via the L1 to L2 retrieval step described above. Also used are meteorological information (especially wind information) as obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis product (Hersbach et al., 2018, 2020). To obtain methane emission estimates the methane column-averaged mole fraction anomalies are converted into vertical column anomalies (in mass per area).

The IUP-UB methane emission estimation algorithm is based on the Cross-Sectional-Flux (CSF) approach. The CSF method is based on integrating the methane column enhancements (in, e.g., tons/km2) times wind speed (in, e.g., km/hour) along several transects (for different distances from the target) perpendicular to wind direction to estimate the methane emission (in, e.g., tons/hour).

CSF algorithms have been developed and used at IUP-UB initially for TROPOMI/S5P methane (Schneising et al., 2020b) and carbon monoxide (Schneising et al., 2020a, 2024) emission estimation and to obtain CO2 emission information from OCO-2 (Reuter et al., 2019) and OCO-3 (Fuentes Andrade et al., 2024). The IUP-UB TROPOMI/S5P methane L4 algorithm described in Schneising et al., 2020b, has been used to obtain emission estimates for large areas such as entire gas and oil fields (e.g., Schneising et al., 2020b) but also to estimate emissions of open pit coal mines (Borchardt et al., 2025).

For MEDUSA related applications a new algorithm has been implemented that can also be applied to other sensors such as PRISMA, EnMAP and EMIT. The latest version of this CSF algorithm is version 1.6 and this version has been used for most of the results produced for the MEDUSA project. This CSF v1.6 algorithm is very similar (but not exactly identical) to the CSF method described in Bösch and Hilker, 2024.

Note that for Parts 1-3 (section 5.1-5.3) we have generated results with v1.3 of our CSF algorithm and the Part 4 (section 5.4) results have been generated with v1.6. Algorithm CSF v1.6 is essentially identical with v1.3 except that 2 minor bugs have been removed (one related to the background correction and a second related to the uncertainty computation). The differences between the results from these two versions are very small (often not even visible in comparison plots), i.e., differences are not relevant for the purpose of the comparisons shown in this document.

The IUP-UB CSF algorithm as used for MEDUSA can be applied to pre-defined target locations. This means that the latitude and longitude of the target location of interest is input for this algorithm. A plume detection algorithm is not used but also not needed for this algorithm. As a consequence, the obtained emission estimate can be zero or even negative (in case of no plume and a noisy methane image).

We rotate the original methane enhancement image (as obtained from the L1-L2 retrieval algorithm) around the target location such that the wind blows to the right (i.e., in the positive x-direction). This has the advantage that the cross-sections along which the methane enhancements are integrated are oriented along the vertical (i.e., y-direction). Only cross-sections located in a box of size 500 m times 500 m are used and quantitatively analysed, where the left side of the box is centred at the target location. To fill data gaps and to obtain an estimate of the background concentration, also the surrounding area is used. Data gaps are filled using Gaussian smoothing. The background correction is based on the median plus a correction for contributions transported by wind into the box.

We compute emission estimates for several cross-sections for different distances from the target location. To minimize the impact of issues related with too short (e.g., complex plume shape close to the source) or too long distances from the target location, we only use emission estimates from cross-sections some small distance away from the target location and using only cross-sections not too far away from the source. For the latter we use a maximum distance corresponding to 3 hours as computed using the wind speed (mostly relevant for TROPOMI due to the large box size of 100 km but not for the 500 m used for HI satellites). The reported emission estimate is the mean of the estimates as obtained for the valid cross-sections.

Wind information is obtained from ECMWF's ERA5 reanalysis product (Hersbach et al., 2018, 2020). We assume that methane is emitted close to the Earth's surface and that on the 500 m scale of interest here the emitted methane is located close to the ground, i.e., stays close to the Earth's surface. We therefore use the 10 m wind (nearest neighbour to the target location) for our CSF method for the HI sensors (for TROPOMI we used a much larger box of size 100 km and the Planetary Boundary Layer (PBL) average of the wind).

The main result of this CSF algorithm is an estimate of the methane emission including its (1-sigma) uncertainty for the given target location and for each satellite overpass, where enough data are available around the target of interest to obtain a robust emission estimate (see below).

The unit of the emission estimate is tons/hour. The emission product files are in CSV format with each line corresponding to a single satellite overpass. Each line contains (at least) the following information: target latitude and longitude, time, estimated emission and its (1-sigma) uncertainty (both in tons/hour), and the quality flag (0 = good).

The uncertainty is estimated by quadratically adding the following terms: (i) The standard deviation of the emission estimates as obtained for the individual cross-sections divided by the square toot of the number of independent measurements (Nindep; computed using the method described in Fuentes Andrade et al., 2024) (ii) a wind related component considering wind variability and a general uncertainty of 0.5 m/s, and (iii) other contributors assuming an additional contribution of 20%.

The quality flag is set to 0, which means “good”, if (i) there are enough data in a box of size 500 x 500 m2 located at the target location (> 90% of area covered), if (ii) Nindep larger than 0, and if (iii) wind speed is not too low (> 1 m/s).

4. Data, comparison structure and methods

In this section, the intercomparison plan for the HI algorithms is described. As mentioned in Sect. 1, we can split it into Parts 1-4, which correspond to section 4.1-4.4, respectively. Detailed plans for intercomparisons of the detection (Sect. 4.1 and Sect. 4.2) and the quantification (Sect. 4.3 and 4.4) of methane emissions with HI data are presented. Note that HI algorithms from some groups use different retrievals for detection and quantification. Note also that IUP-UB will not participate in the plume identification and attribution to source intercomparison. Instead, they will focus on methane image comparisons and emission quantification. Source location information is input for IUP-UB and plume delineation is not needed for their CSF-based emission quantification algorithm.

4.1. Plume identification and attribution to source

There are several aspects of methane emission detection that need to be analysed. An EnMAP image capturing an area in the Shanxi coal mines (China) and an EMIT acquisition of an O&G region in Turkmenistan (Table 1) have been selected to assess these aspects. The EnMAP scene exhibits a relatively dark albedo and a heterogeneous surface, while the EMIT one is bright and homogeneous. Both acquisitions contain numerous plumes in a wide range of emission rate values. As a result, these images emerge as promising candidates for studying emission identification and the associated source attribution for each group's algorithm. Note that we account for the whole extension of the scenes.

Some issues can arise after the examination of these images such as the plume segmentation for detection purposes (it can be different for quantification), the criteria followed to decide whether a specific source is emitting, and the criteria to distinguish between plumes that are spatially close. Differences in algorithms regarding these points can cause discrepancies in detection. In addition, identifying plumes such as small emissions or those with atypical plume shapes can be challenging with automated methodologies, instead of with visual inspection, which allows a more flexible detection. Differences regarding these methodologies should also be analysed.

Mission	Timestamp (YYYYMMDDThhmmss)	Image central latitude	Image central longitude	Methane Sector	Location	Characteristics
EnMAP	20230330T033620	36.1380	113.1030	Coal	China (Shanxi)	dark, heterogeneous, numerous plumes, mixed plumes
EMIT	20220827T060753	36.2890	61.7605	O&G	Turkmenistan	bright, homogeneous, numerous plumes, mixed plumes

Table 1. *Information about the data used for the 'Plume identification and attribution to source' intercomparison.*

Group outputs:

- A CSV file with the latitude and longitude coordinates of every emitting source.
- Delineations of every detected plume and its association to a specific source. Plume masks will be provided in NetCDF format and the attribution to source will be included in the CSV file next to the emitter coordinates.

Intercomparison:

A potential emitter (PE) detection matrix will be created (as in the example shown in Fig. 1) and the number of detected plume emitters will be compared and discussed. Bar charts will also be used for comparison. There will be a special focus on the use of automated algorithms and the potential false positives due to retrieval artifacts. In addition, plume delineations and the attribution to source will also be compared and discussed. As for the source attribution comparison, we will calculate the distance among group locations in terms of distance (in meters) and number of pixels using QGIS. Finally, some of the plume masks from different groups will be overlapped to examine the mask extension (see example in Fig. 2) and the decisions made on mixed plumes from different emitters.

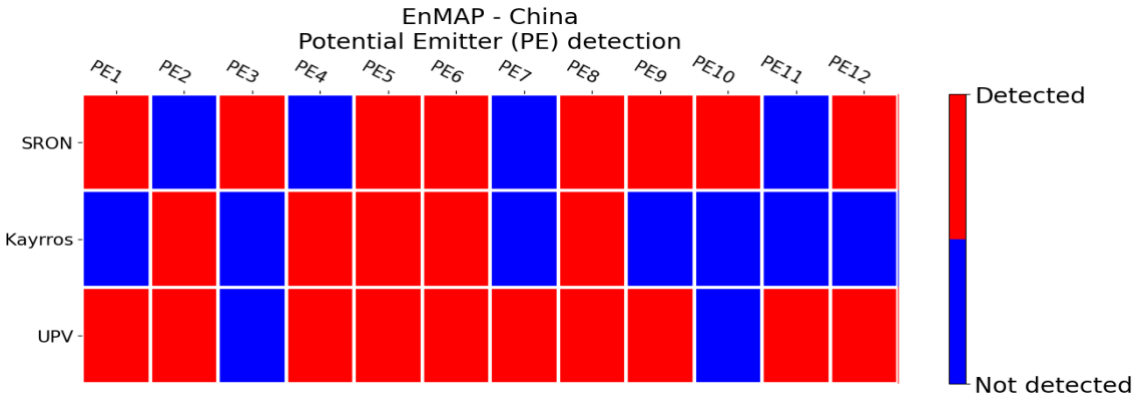


Figure 1. *Example of a potential emitter detection matrix where the group names and the potential emitter identifiers are in the y- and x- axis, respectively. Detections are in red and not detections are in blue.*

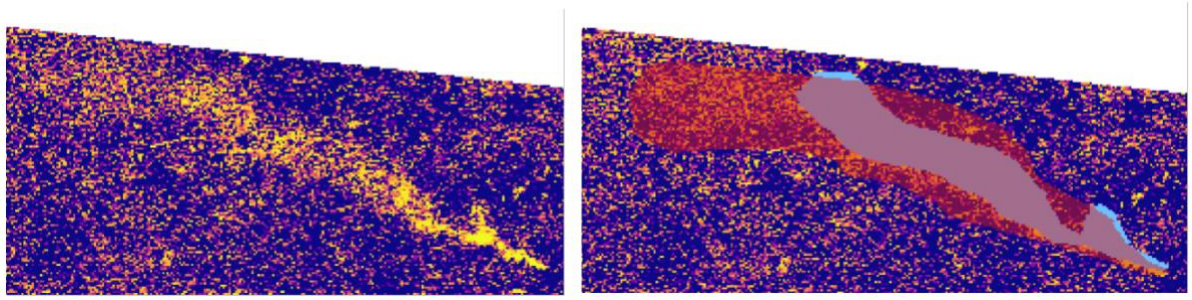


Figure 2. *Methane retrieval showing a plume without (left panel) and with (right panel) two different overlapped plume delineations (in red and in blue).*

4.2. Assessment of methane concentration retrievals used for detection

Methane emission detection is based on methane concentration maps, where the plume can be detected by automated methodologies or visual inspection. In any case, the detection capabilities will be mostly determined by the retrieval random noise and the number of retrieval artifacts that disturb the identification of real plumes (Roger et al., 2024). These originate from several instrument and surface features, but also from the detection algorithm performance.

Methane retrievals used for detection from each group algorithm will be analysed to assess the impact of random noise and retrieval artifacts on detection. Plume-free data is selected to avoid the influence of methane emissions in this assessment.

9 images capturing areas with different levels of radiance and heterogeneity are used to study the influence of these factors on detection (see Table 2). A selection of 100 x 100 pixel subset is made because is considered representative enough for the study and also in order to reduce the processing time.

A retrieval histogram along to the 1-standard deviation and mean values can be used to study the emission detection capability (Fig. 3). 1-standard deviation can be understood as the retrieval precision, while the mean deviation from the zero value can be generally seen as the presence of retrieval artifacts.

Mission	Timestamp (YYYYMMDDThhmmss)	Image central latitude	Image central longitude	Location	Surface characteristics	Subset corners (xi,yi ,xe,ye)
EnMAP	20230830T014338	-28.8953	136.3137	Australia (Anna Creek)	bright, heterogeneous	100,800, 200,900
PRISMA	20231103T010502	-28.7314	136.28090	Australia (Anna Creek)	bright, heterogeneous	250,350, 790,890
EMIT	20231124T020552	-29.1747	136.1519	Australia (Anna Creek)	bright, heterogeneous	700,700, 800,800
EnMAP	20240628T053313	44.8271	88.9972	China (Xinjiang)	bright, homogeneous	310,900, 410,1000
PRISMA	20220603T050600	44.853	88.7801	China (Xinjiang)	bright, homogeneous	0,0, 100,100
EMIT	20220815T025827	45.0731	88.4961	China (Xinjiang)	bright, homogeneous	50,150, 150,250
EnMAP	20230426T114318	40.3969	-3.5773	Spain (Madrid)	dark, heterogeneous	50,340, 150,440
PRISMA	20211117T111916	40.3373	-3.5817	Spain (Madrid)	dark, heterogeneous	50,150, 150,250
EMIT	20231007T105048	40.8483	-3.3539	Spain (Madrid)	dark, heterogeneous	305,405, 175,275

Table 2. *Information about the data used for the 'Assessment of methane concentration retrievals used for detection' intercomparison. Note that xi, yi, xe, ye are the pixel indexes from the L1 data that define the subset selection.*

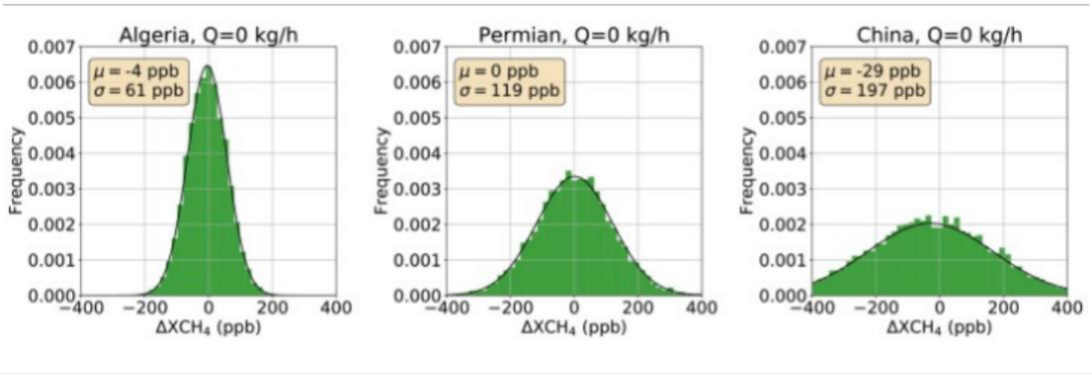


Figure 3. *Methane retrieval histograms of 100 x 100 pixel subsets from plume-free areas. 1-standard deviation (σ) and mean (μ) values are also shown (from Guanter et al., 2021).*

Group outputs:

- As in Fig. 3, histograms from the retrievals used for detection along with 1-standard deviation and mean values in PNG format.

Intercomparison:

Tables showing the 1-standard deviation and mean values from the different data will be created. On the other hand, group retrievals will be shown to illustrate differences. The associated discussion will be linked to the group HI algorithms and the data surface features.

4.3. General methane emission quantification assessment

We have carefully selected 12 acquisitions containing methane emissions from the EnMAP, PRISMA, and EMIT instruments that are representative of a wide range of scenarios (Table 3). Among these scenarios we find images with different levels of radiance and heterogeneity, and even with the presence of water bodies in images acquired from land areas. Note that, for instance, the presence of water bodies, such as lakes, in land scenes could impair the calculation of robust statistics using the matched filter method. We mostly find methane emissions from point sources, although we have included one from an area-source. Moreover, we include emissions coming from different industries, such as the oil and gas, coal mines and waste. Two additional plumes that were artificially generated for validation purposes in two different controlled release experiments (Sherwin et al, 2023; Sherwin et al., 2024) are also included. Finally, we also account for plumes with different length, intensity, and shape. In this section, we will compare the flux rate estimates from each group's algorithm for all of these emissions. In this manner, we will be able to detect general biases between groups and to identify particular situations that cause discrepancy.

Mission	YYYYMMDDThhmmss	Plume source latitude	Plume source longitude	Methane Sector	Location	Characteristics
EnMAP	20221002T074833	38.6768	54.3035	O&G	Turkmenistan	bright, homogeneous, long plume
EnMAP	20221116T184050	32.82182	-111.785773	Validation	US (Arizona)	bright, heterogeneous
EnMAP	20231027T075753	45.3324	52.373	O&G	Kazakhstan (Karaturun East oil field)	strong emission, water presence, bright, heterogeneous
EnMAP	20240111T055606	28.62515	77.32809	Waste	India (Delhi)	area source, dark, heterogeneous
PRISMA	20200703T071439	38.55747	54.20049	O&G	Turkmenistan	bright, homogeneous
PRISMA	20200421T173935	32.057	-103.533	O&G	Permian (US)	bright, heterogeneous
PRISMA	20210206T031837	36.2342	112.949	Coal	China (Shanxi)	dark, heterogeneous
PRISMA	20211027T182628	33.63064	-114.48915	Validation	US (Arizona)	atyplical plume, dark, heterogeneous
EMIT	20230223T063033	39.97369	61.68137	O&G	Turkmenistan	bright, homogeneous, long plume
EMIT	20230624T175855	32.4513	-101.4629	O&G	Permian (US)	bright, heterogeneous
EMIT	20240415T062152	63.2305	112.9502	Coal	China (Shanxi)	dark, heterogeneous
EMIT	20240405T170129	38.55747	54.20049	O&G	US (Kansas)	strong emission, dark, heterogeneous, cirrus presence

Table 3. *Information about the data used for the 'General quantification assessment' intercomparison.*

Group outputs:

- CSV file with the flux rate estimates and the used wind speed values (including uncertainties).

Intercomparison:

Bar charts comparing the different group flux rate estimations from the different emissions (E) as in the artificial example shown in Fig. 4 will be created. We can observe a systematic overestimation and underestimation in reference to the rest of the groups of the blue and red groups, respectively. Thus, this kind of chart would be able to provide general tendencies between groups. We will also show the wind speed values used by each group to assess their variability and their impact in the quantification.

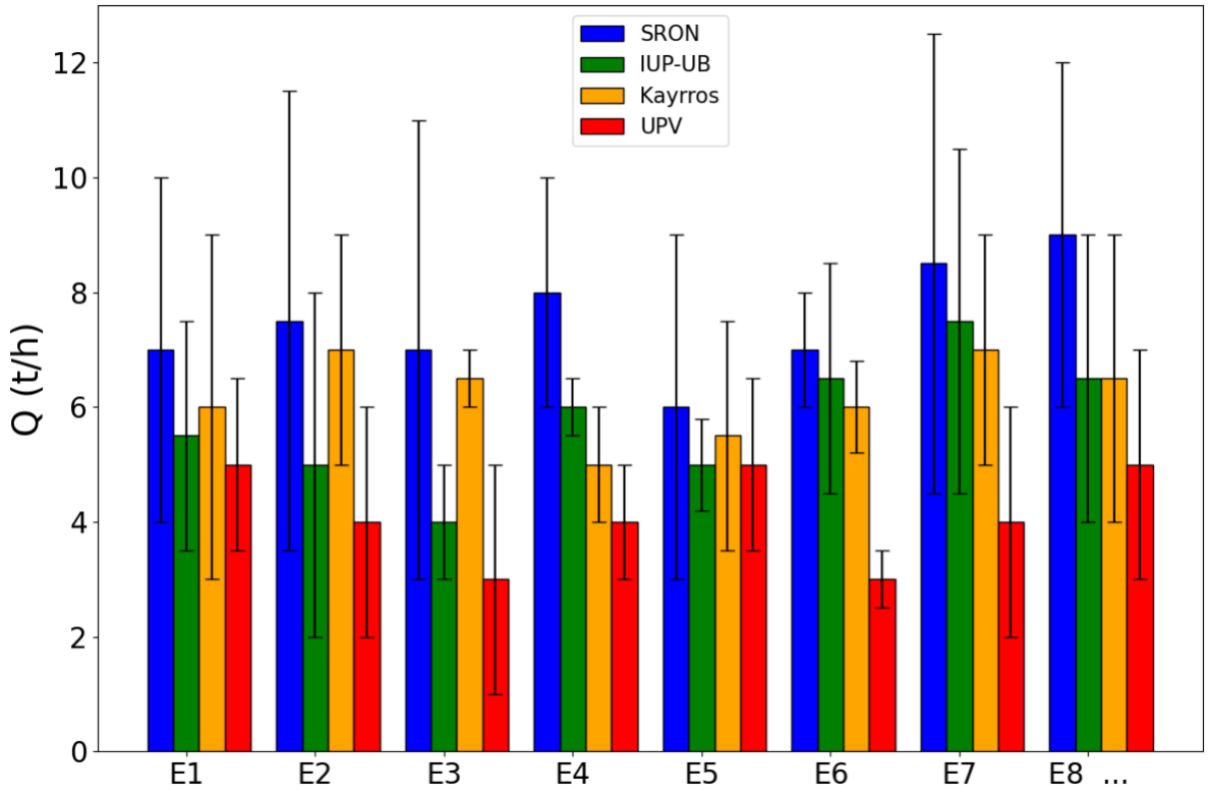


Figure 4 *Example of bar charts comparing the flux rate of several emissions (E) from the SRON (blue), IUP-UB (green), Kayrros (orange), and UPV (red).*

4.4. Detailed methane emission quantification assessment

4.4.1. Methane retrievals used for quantification

The flux rate estimation from a methane emission involves a relatively long processing chain with numerous intermediate steps. Methane retrievals used for this purpose constitute a significant portion of these steps and provide a baseline for applying quantification methods. Therefore, a thorough analysis of this product is necessary to assess its impact on the final flux rate estimation.

Biases between the retrievals from the different algorithms can be found because of several reasons. The use of different spectral windows, cloud and water masking in the radiance data, the use of an iterative step to respect the matched-filter related sparsity assumption, an employment of different Jacobians, and how the different algorithms deal with strong emissions and saturation are some of these factors. Studying retrieval biases will allow us to better understand the influence of these points in the retrieval and therefore in the final flux rate estimation. We will select several emissions from Table 3 to carry out this analysis. This selection will be made in Sect. 5.4.1, after finishing the intercomparison and discussion from Sect. 5.3. These emissions will be selected to include all the instruments and preserve a certain degree of variability in emission and scene features

Group outputs:

Methane retrieval (used for quantification) subsets in ppm units containing the plume pixels in NetCDF format. An example of the NetCDF structure file is shown in Fig. 5.

```
<class 'netCDF4._netCDF4.Dataset'>
root group (NETCDF4 data model, file format HDF5):
    description: Example: NetCDF file that includes quantification retrieval in parts-per-million (ppm) units
    dimensions(sizes): row(250), col(900)
    variables(dimensions): float32 Delta_XCH4(row, col)
    groups:
```

Figure 5. *Example structure of a NetCDF file containing a non-georeferenced methane concentration enhancement map used for quantification, with 250 rows and 900 columns, in ppm units.*

Intercomparison:

Retrievals will be compared similarly to the example matrix shown in Fig. 6, where the diagonal elements are the original group retrievals, and the not diagonal elements are the retrieval difference and the associated scatter plot. In this example, we can see that the Group1 algorithm provides lower plume values than Group2 close to the source, where methane enhancement values are usually high. This can be an indicator of saturation of the Group1 algorithm in reference to that of the Group2. We can verify this by addressing relevant aspects of the group algorithms for HI (see Sect. 3). We will obtain a considerable

amount of these figures because of the high dimensionality coming from the large number of groups. Therefore, instead of showing all these figures, we will keep the most important plots for the analysis. Note that there will be surface elements that could have an impact on plume values even though they do not appear in the retrieval subset. This influence will also be considered.

At this point, differences among group retrieval pixel values might not be understood yet. Depending on the results, we can consider carrying out a more specific analysis.

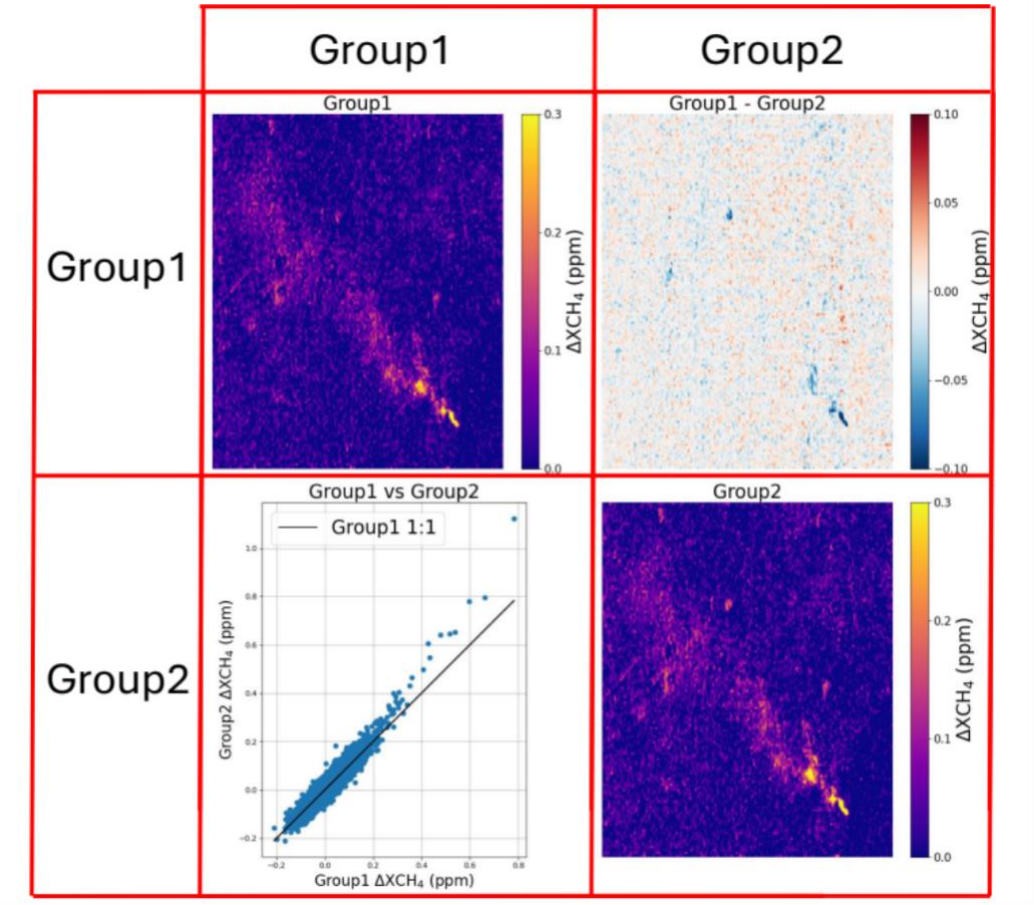


Figure 6. *Example matrix, where the diagonal elements are the original group retrievals containing a plume, and the not diagonal elements are the retrieval difference (upper-right) and the associated scatter plot (bottom-left). Retrievals show methane concentration enhancement values in parts-per-million (ppm).*

4.4.2. Comprehensive quantification assessment

Due to the relatively long processing chain involving numerous steps that can alter the final flux rate estimation, we will compare estimations accounting for the influence of some of the intermediate products. We consider 4 cases:

A) No common intermediate products.

B) Common wind speed value, i.e. wind speed at 10-m above surface (U_{10}).

C) Common methane concentration enhancement map.

D) Common wind speed value (U_{10}) and methane concentration enhancement map.

In this manner, we can individually assess the impact of the quantification methods, retrieval methods, or the selection of the wind speed values on the flux rate estimation. Then, if a bias arises in scenario A), we can identify the most significant contributing factors by examining the remaining cases.

We will use the same data as in the 'Methane retrievals used for quantification' section, so we can also account for the previous extracted conclusions. Thus, a comprehensive analysis of key processing steps for flux rate emission estimation will be carried out, which will allow us to better understand flux rate biases between group algorithms.

For case B) and D), a common U_{10} value of 3 m/s with zero uncertainty is considered for every case. Regarding case C) and D), common methane retrievals will be provided by UPV. However, these retrievals are not processed according to the UPV algorithm described in Sect. 3 as previously proposed in the AIP. Instead, UPV will share retrievals that are obtained as described in Guanter et al., (2021). This will allow us to also evaluate a change in reference to the original UPV algorithm, which can be beneficial for the analysis. Moreover, in case we need to understand differences in reference to the shared retrievals, it would be easier since the algorithm involved is less complex than the one used by UPV.

Group outputs:

- CSV file with the flux rate estimates and the used wind speed value along to their uncertainties for the 4 cases.

Intercomparison:

Bar charts comparing the different group flux rate estimations per case as in the artificial example shown in Fig. 7 will be created. We can observe that flux rate estimations for every group are very similar in case D), which points to the wind speed and methane retrievals as the discrepancy sources. If values from case D) had shown a higher discrepancy, the

quantification method could likely be identified as the main source of this difference. Thus, similar discussions will be made to identify main differences between algorithms. We will also show the wind speed values used by each group to assess their variability and their impact in the quantification. If these figures are not sufficient for the analysis, we will consider creating a matrix that shows the ratio of the groups flux rates, using their distance from 1 as a measure of discrepancy.

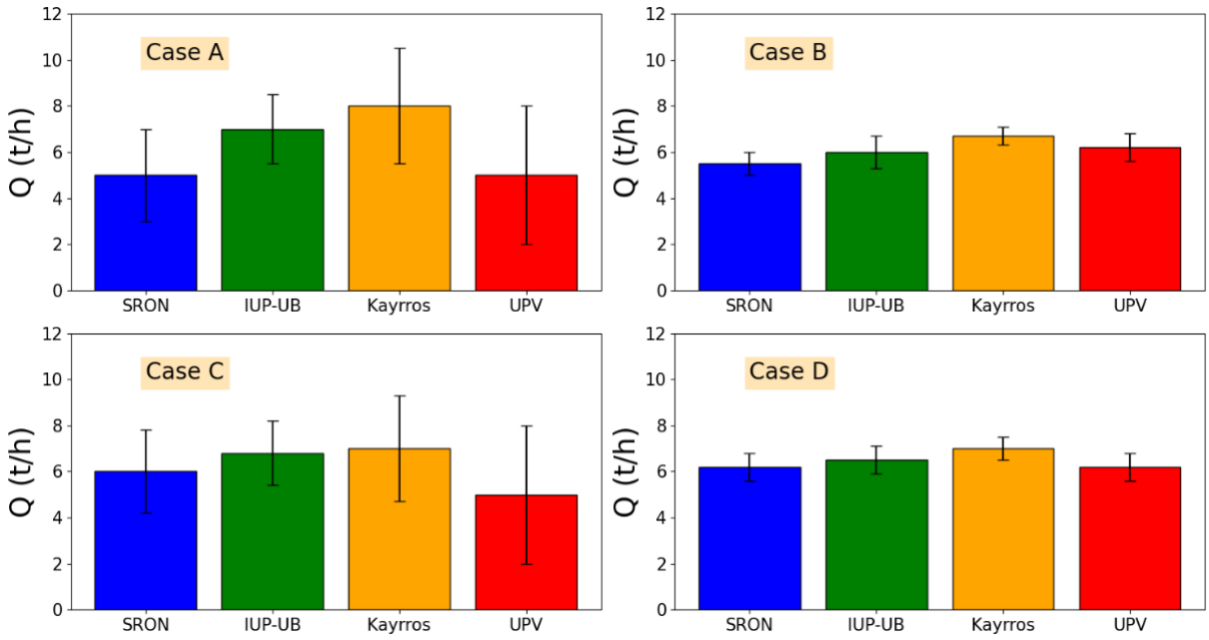


Figure 7: *Example of bar charts comparing the flux rate estimations from the SRON (blue), IUP-UB (green), Kayrros (orange), and UPV (red) in the case A (upper-left), B (upper-right), C (bottom-left), and D (bottom-right).*

5. Results and discussion

5.1. Plume identification and attribution to source

We use an EnMAP and an EMIT acquisition (see Fig. 8) to assess the emission identification and the attribution to a source. Several algorithm choices are important in the examination of these images such as the plume segmentation for detection purposes, the criteria followed to decide whether a specific source is emitting, and the criteria to distinguish between plumes that are spatially close. Differences in algorithms regarding these points can cause discrepancies in detection. In addition, identifying plumes from small emissions or with atypical plume shapes can be challenging with automated methodologies, instead of with visual inspection, which allows a more flexible detection. IUP-UB did not participate in this part, as they had not developed a detection algorithm at the time of this exercise.

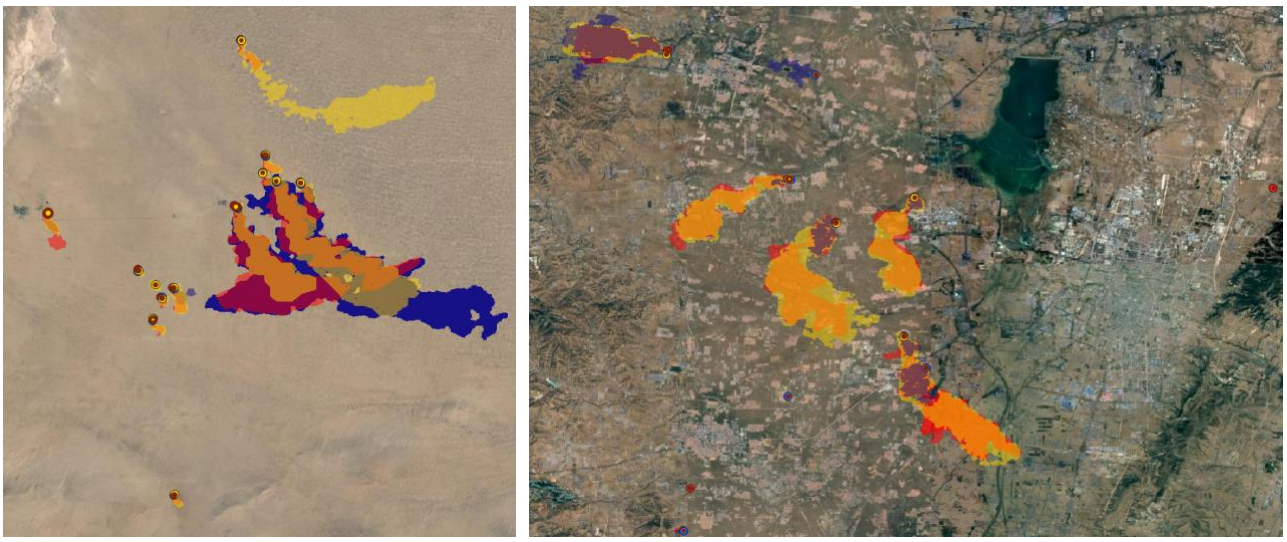


Figure 8. *Potential emitter locations (points) and plume delineations of the SRON (blue), Kayrros (yellow), and UPV (red) groups for the EMIT scene in Turkmenistan (left) and the EnMAP scene in China (right).*

Potential Emiter identification

For the EMIT acquisition, there is strong agreement among groups regarding the Potential Emitters (PE) selection, which are those locations that are identified as sources that are emitting a methane plume at acquisition time. As shown in Fig. 9, SRON detected 11 PE, UPV 12, and Kayrros 13. The high brightness and homogeneity of the EMIT scene allow strong performance of all algorithms, generally providing clear plumes. The only PE that was detected by just one group was PE9, which was identified by Kayrros. This emitter was also detected in Thorpe (2023), which adds credibility as an emitting source. Disagreement among groups comes only in the case of the smallest plumes, such as in PE9 and PE13. While SRON and UPV use a fully (or semi) supervised approach for detection, Kayrros uses a deep learning-based model followed by a manual review. Under this specific scenario, the latter seems to provide a more comprehensive list of plume candidates, which outnumbers those obtained by only using a fully manual approach. However, this conclusion cannot be generalized since we would need a significantly higher number of scenes to test this point.

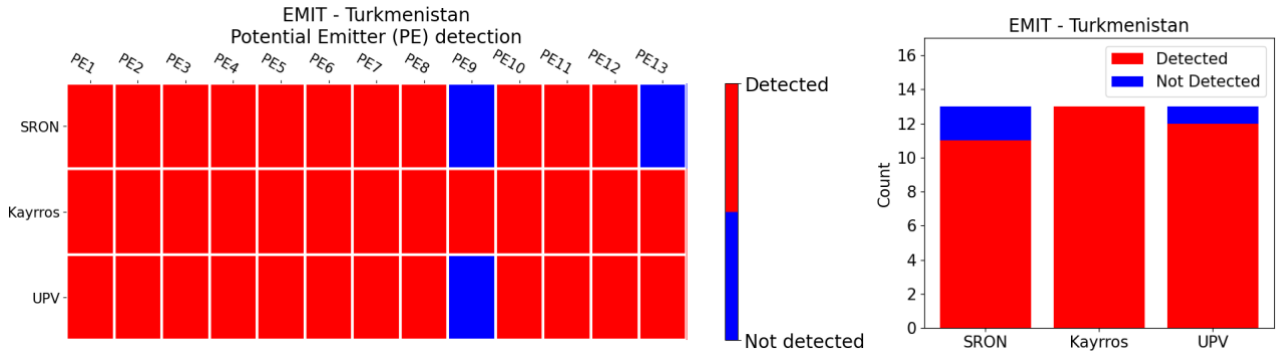


Figure 9. *On the left panel, Potential Emitter (PE) detection matrix showing the detected (red) and not-detected (blue) PE (columns) by each group (rows) based on the EMIT scene capturing a Turkmenistan area. On the right panel, a bar chart showing the count of detected and not detected PE (y-axis) for each group (x-axis) for the same scene.*

For the EnMAP instrument, the three groups use a fully manual approach for detection. The EnMAP scene led to higher disagreement among groups. As shown in Fig. 10, SRON detected 9 PE, UPV 10, and Kayrros 5. The higher heterogeneity and the lower brightness of the scene result in retrievals with more noise. Therefore, detection is more challenging, and disagreement is more likely to occur. The Kayrros team uses a more conservative approach regarding manual detection, since only the stronger and clearer plumes were identified. Meanwhile, UPV and SRON featured more detections in these more challenging circumstances. However, it is worth noting that a detection of a PE does not always imply a detection of a true plume. We searched on the Carbon Mapper portal (https://data.carbonmapper.org/) to check whether those PE identified by only one group show emissions in the portal records. We found that PE3 and PE10, which were only detected by SRON, registered emissions. Similarly, PE11, which was only identified by UPV, also registered emissions in the portal. On the other hand, PE7, which was also only detected by UPV, had no emissions. After a more careful inspection, UPV suggested that the initially detected plume from PE7 was probably a retrieval artifact originating from surface-related infrastructure. This illustrates that in any detection approach, attention should be paid to false positives.

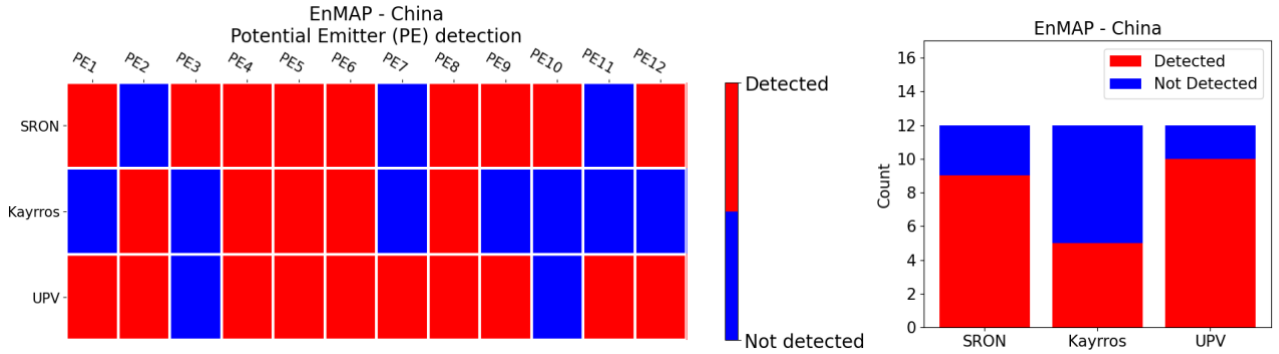


Figure 10. *On the left panel, PE detection matrix showing the detected (red) and not-detected (blue) PE (columns) by each group (rows) based on the EnMAP scene capturing a China area. On the right panel, a bar chart showing the count of detected and not detected PE (y-axis) for each group (x-axis) for the same scene.*

Plume delineation

We also assessed differences in plume delineation from those cases where groups shared the same PE. Several cases are illustrated to analyze these differences.

Case 1

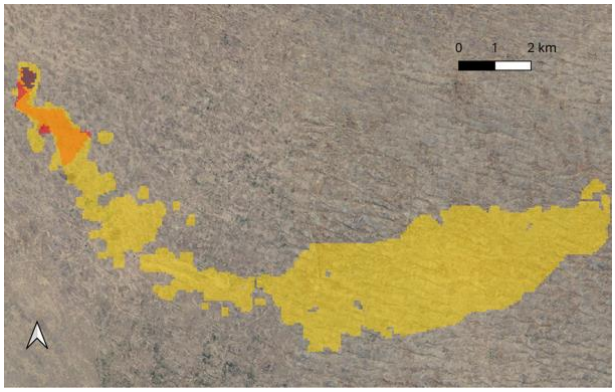


Figure 11. *Plume delineations in the EMIT scene of the SRON (blue; shown underneath the other colored masks), Kayrros (yellow), and UPV (red) groups for Case 1.*

In this case, we find large discrepancy among the plume delineations, where Kayrros shows a Plume Length (PL) remarkably higher than the rest of the groups. For the EMIT scene, Kayrros uses a deep learning model followed by a manual review. This approach can generate delineations that combine not connected clusters, which helps to capture a higher extension of the plume in those cases where the methane enhancement pixels are distant from each other. This happens in some plume cases due to factors such as turbulence. Moreover, UPV reports a PL significantly higher than SRON's, but still substantially lower than that reported by Kayrros.

Case 2

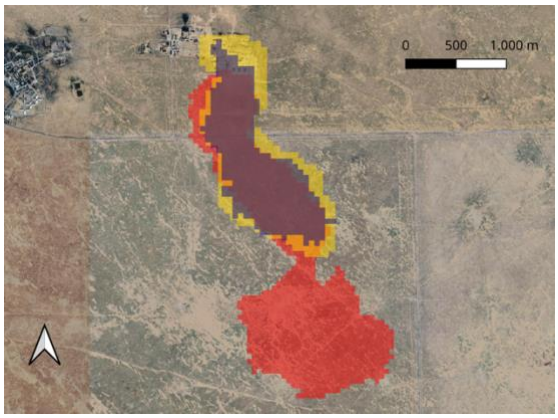


Figure 12. *Plume delineations in the EMIT scene of the SRON (blue), Kayrros (yellow), and UPV (red) groups for Case 2.*

PL from SRON and Kayrros are similar, while UPV considers a higher extension beyond the common plume. UPV plume delineation is mainly based on the retrieval visual inspection, which makes the mask selection highly dependent on the user criteria. This leads to a higher ability to identify relatively low methane enhancement values, but there is also the risk of confusing the potential methane enhancement with retrieval background noise.

Case 3

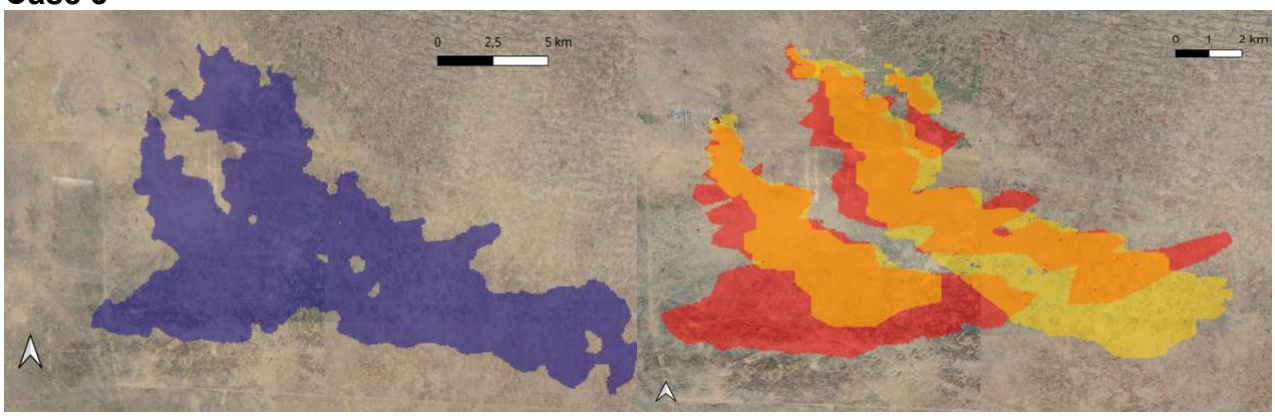


Figure 13. *On the left panel, plume delineations in the EMIT scene of the SRON (blue) group for Case 3. On the right panel, Kayrros (yellow), and UPV (red) plume delineations for the same case.*

All the teams identified the same PE from this plume complex. Here the plumes are spatially close and even overlapped. SRON used the masking of the whole complex to address the combined emissions coming from the different PE. This is a more conservative approach, since it is challenging to properly distinguish among mixed plumes. On the other hand, Kayrros and UPV delineated individual plumes, which can be useful for quantification purposes.

Case 4

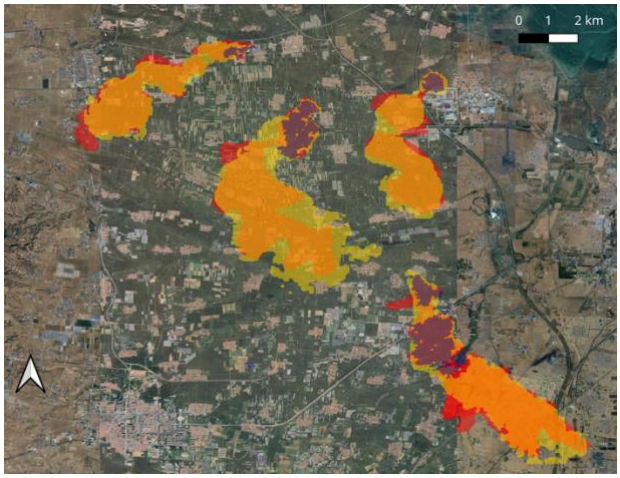


Figure 14. *Plume delineations in the EnMAP scene of the SRON (blue), Kayrros (yellow), and UPV (red) groups for Case 4.*

PL from UPV and Kayrros are similar, while SRON's is more restrictive. For the EnMAP case, Kayrros uses a fully manual approach for plume delineation, similarly to UPV. The visual inspection criteria from these 2 groups exhibit good agreement, although some differences can be found at the plume edges. SRON uses a masking retrieval threshold of 2 to 3 standard deviations of background methane values. Since this EnMAP scene captures a remarkably heterogeneous area, we obtain a relatively poor retrieval precision. This leads to a more restrictive mask from the watershedding method that aligns with the shorter PL. Similarly in Case 1, even though the EMIT scene shows a better retrieval precision, we also observed that UPV and Kayrros PL were higher than SRON's. Therefore, a masking based on this threshold approach seems to generally lead to shorter PL.

Case 5

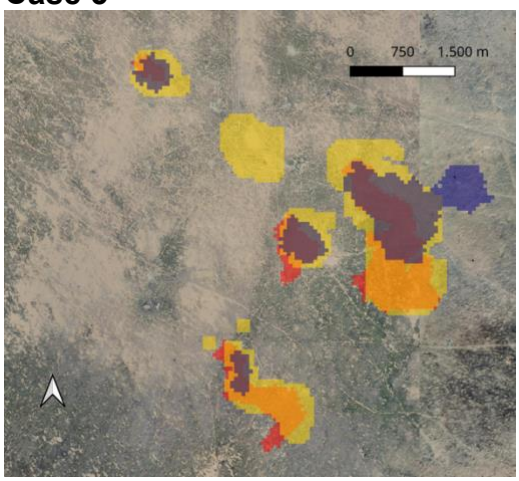


Figure 15. *Plume delineations in the EMIT scene of the SRON (blue), Kayrros (yellow), and UPV (red) groups for Case 4.*

In this case, we observe several emissions, generally small and even blob-like. We identify substantial differences in the plume shapes regarding the groups' delineations. Due to the typical low enhancement values from these emissions, it is difficult to distinguish between background noise and methane, which aligns with this discrepancy. Here, we also identify SRON as the group with generally shorter PL, which matches the more conservative approach based on retrieval precision thresholds.

Altogether, we can extract the following conclusions after analyzing these 5 cases:

- Kayrros allows plume delineation for not connected methane enhancements when using a deep learning model approach.
- SRON generally performs a more restrictive plume delineation and follows a more conservative approach when masking overlapping plumes.
- UPV strongly follows a visual inspection-based masking, which can lead to background noise pixels within the plumes.
- Depending on the emissions, there can be remarkable differences.

Additionally, the plume delineation performed by these groups is also used to extract the methane emission flux rate. The quantification methods from each group are highly dependent on how they specifically characterize the involved parameters. As for the IME quantification, the effective wind speed parameter and the plume length are carefully selected to be consistent with the plume delineation. Therefore, plume delineation differences among groups should not be a source of emission flux rate discrepancy as long as the group methods are consistent with their plume masks and use a wind speed calibration that was obtained using the same approach. However, UPV and Kayrros use the same effective wind speed calibration and the same criteria to determine the plume length when using the EnMAP instrument. In case 4, we observed there were no major differences among the plume delineations from these two groups, which provides consistency for quantification in this specific case.

Attribution to source

In Fig. 16, we have compared the relative distance in terms of meters (left) and in terms of number of pixels (right) among the locations assigned by each group for those PE that were detected by at least 2 groups. In terms of meters, we show the distribution of distances between pairs of group PE locations (black) using the mean (dot) and the standard deviation (bars). The best agreement is found for UPV-Kayrros (90 m) and the highest discrepancy is found in SRON-UPV (170 m). The analysis is based on only ~15 points, which means there is limited generalizability, as reflected by the large error bars. Therefore, in the case of individual instrument data, we used the median of the distribution. Moreover, we see that the tendency is similar for both the EnMAP (green) and EMIT (red) instruments. However, absolute distances are greater for EMIT due to its higher GSD and this is why we also show the distance in terms of the number of pixels. The observed shifts might come from different orthorectification approaches. Kayrros, SRON, and UPV combine orthorectified retrievals with high resolution imagery to pinpoint the source. However, while Kayrros and SRON always pinpoint the potential source location within their plume masks, UPV can pinpoint the potential source outside the plume mask if they find a very clear piece of infrastructure that can be related to the source. Therefore, discrepancies can arise due to differences in each group's criteria. Altogether, we generally find differences of only approximately 1-3 pixels, which is translated into 30 m – 90 m for EnMAP, and 60 m – 180 m for EMIT.

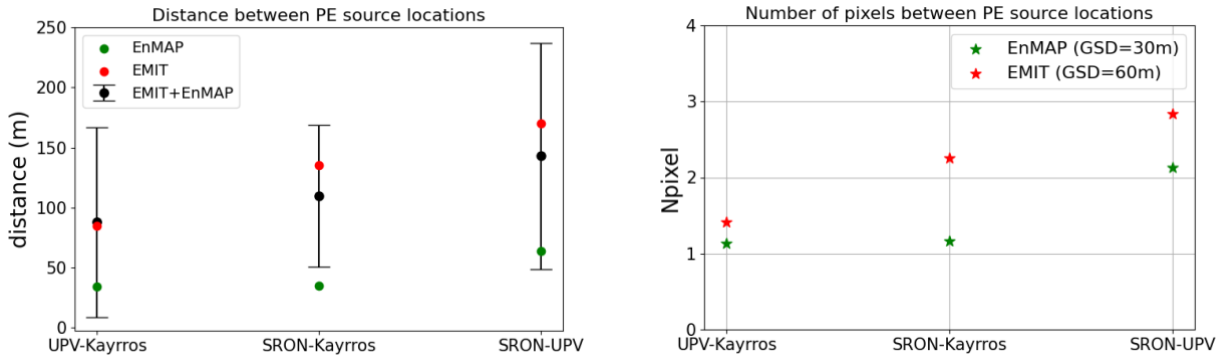


Figure 16. *On the left panel, mean (black dots) and standard deviation (error bar) of the distribution of distances in meters of PE locations among groups. The median of the distribution for individual instruments is shown for EMIT (red) and EnMAP (green). On the right panel, the median of the distribution in terms of numbers of pixels (Npixel) for individual instruments is shown with the same color code.*

5.2. Assessment of methane concentration retrievals used for detection

Several groups use different retrievals for quantification and detection. Regarding the latter, random noise and retrieval artifacts have a very strong impact on plume identification (Roger, 2024). In this context, detection retrievals of plume-free data from the different groups will be analyzed to assess their performance. PRISMA, EnMAP, and EMIT acquisitions from three different areas have been used for this purpose. First, we look at a smaller group of cases to extract some insights and then we will analyze all results together.

EnMAP – Australia (Anna Creek)

Fig. 17 shows the retrievals of the 4 groups over a selected region in Australia from the EnMAP acquisition. The IUP-UB algorithm uses a classical matched filter in the 2300 nm methane absorption window, while Kayrros uses a wider window that also includes the 1700 nm absorption window. We find a retrieval artifact in the upper part of the retrieval, which seems stronger in Kayrros' retrieval. The extended window algorithm increases the related score, i.e. retrieval enhancement level, in reference to the narrow window version, probably due to the surface related structure having similar features to those of the methane absorption in the 1700 nm window. This also happens in other parts of the subset, but there is also the reduction of enhancement values in some other parts. On the other hand, UPV and SRON use the Combo-MF method (Roger, 2024), which is a matched filter flavor that has proven to be favorable for plume detection. It keeps the attenuation of background retrieval values from the extended window and penalizes those enhancement values that are higher in the extended window retrieval in reference to the one in 2300 nm, while approximately preserving the methane-emission-related pixels. We can observe that Combo-MF values generally exhibit relatively low score in those pixels where Kayrros is lower than IUP-UB and also in those pixels where Kayrros is higher than IUP-UB. The Combo-MF penalization leads to a higher number of negative values in the retrieval distribution, which results in a negative distribution mean (see Fig. 18). On the other hand, the positive means from the Kayrros and IUP-UB retrievals occur due to the presence of positive retrieval artifacts. Note that even if UPV and SRON use the Combo-MF method, they differ in other aspects such as the exact spectral coverage of the spectrum, i.e. SRON uses a 1300 nm – 2500 nm window, while UPV uses a 1000 nm – 2500 nm spectral range.

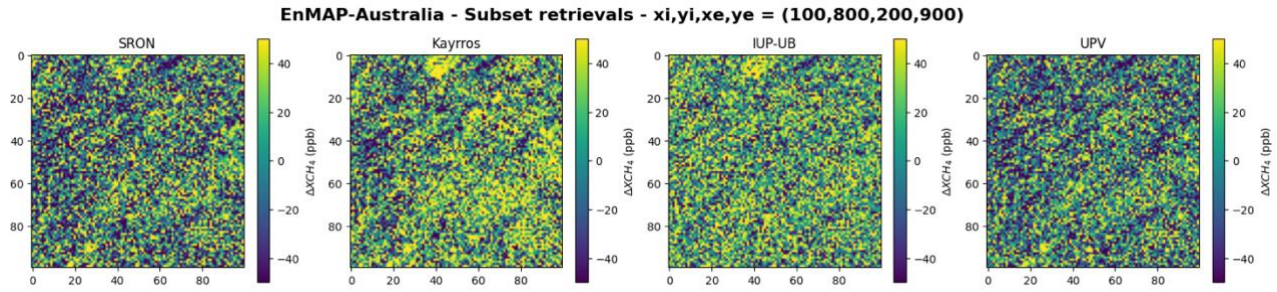


Figure 17. *Methane retrievals from the selected subset of the EnMAP acquisition capturing an Anna Creek (Australia) area of the (starting from the left) SRON, Kayrros, IUP-UB, and UPV groups.*

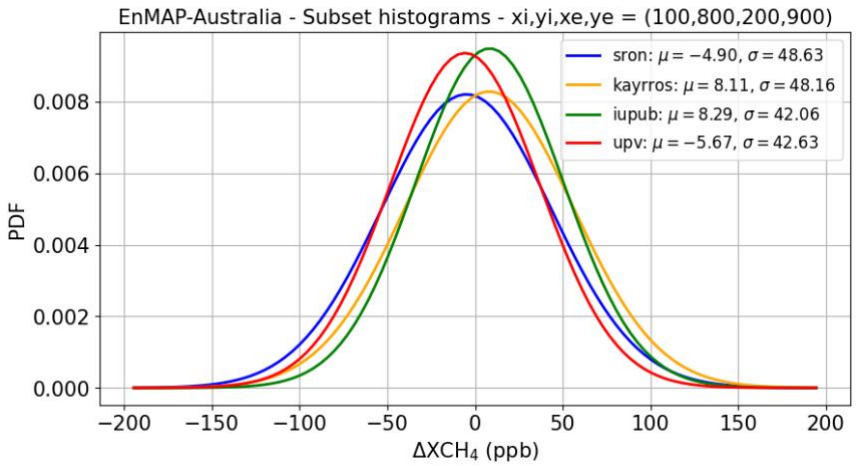


***Figure 18**. Histograms showing the methane concentration enhancement distribution of the SRON (blue), Kayrros (yellow), IUP-UB (green), and UPV (red) retrievals from the selected subset of the EnMAP acquisition capturing an Anna Creek (Australia) area.*

As shown in Fig. 19, the apparent large difference between IUP-UB and UPV is mainly due to a small offset between the 2 images. If an offset is added to the IUP-UB image such that both images have the same mean (-5.7 ppb) then both images agree quite well (Fig. 19 - bottom). Since UPV uses the Combo-MF, the higher frequency of negative values is due to the penalization of background pixels, which leads to the shift of the distribution compared to matched filter-derived retrievals that use the 2300 nm absorption window (as IUP-UB).

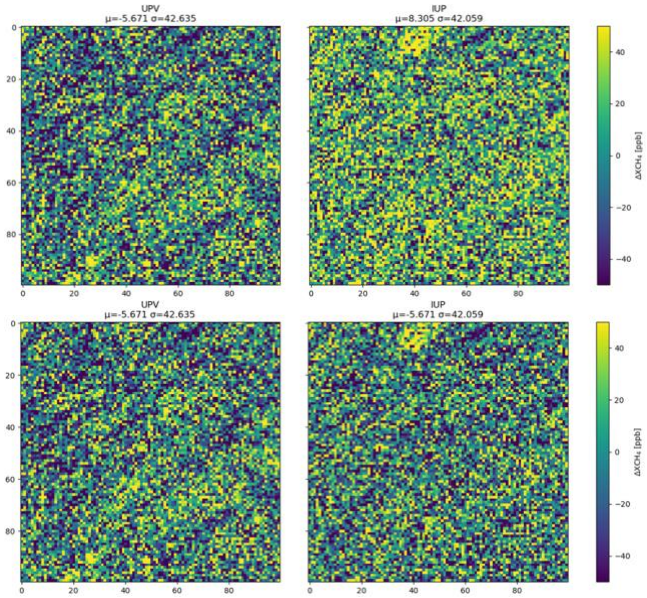


***Figure 19:** Similar as Fig. 17 but limited to IUP-UB (right images) and UPV (left images). Top: As in Fig. 17. Botton: As top, but with small offset added to IUP-UB.*

EnMAP –Shanxi (China)

In Fig. 20, we also observe a strong retrieval artifact in IUP-UB appearing in the upper part of the subset. In this case, we can observe a remarkable artifact attenuation in the Kayrros retrieval, which is translated in the wider window retrieval being able to better distinguish between the artifact and methane. This is probably due to the surface related structure not having similar features to methane in the 1700 nm window, which allows for a better discrimination. However, below the artifact the IUP-UB shows very low values, which are increased in the Kayrros retrieval. Again, we observe that Combo-MF from SRON and UPV can generally better attenuate the appearance of retrieval artifacts. As a result, we find overall lower scores in these retrievals. Regardless of the attenuation or amplification of the retrieval artifact in the wide-window retrieval, Combo-MF appears to improve plume detection capabilities.

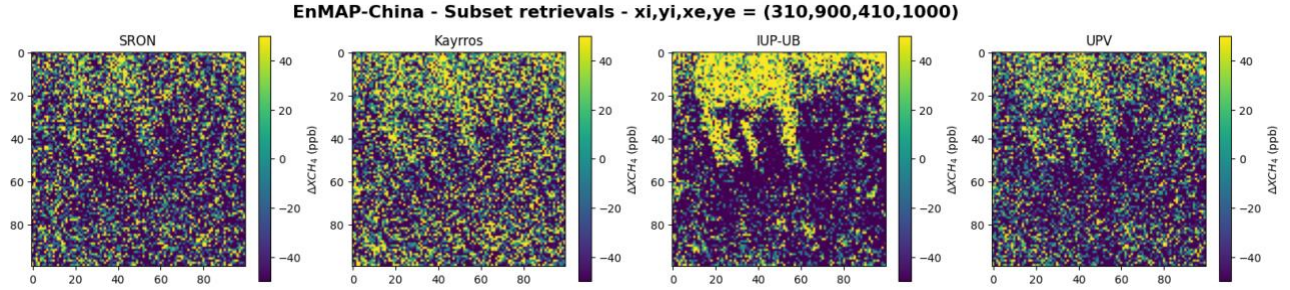


Figure 20. *Methane retrievals from the selected subset of the EnMAP acquisition capturing a Shanxi (China) area of the (starting from the left) SRON, Kayrros, IUP-UB, and UPV groups.*

Based on the initial IUP-UB results shown in Fig. 20, IUP-UB has carried out further analysis. The results are shown in Fig. 21. In this context, a bug related to cloud masking has been identified for the IUP-UB results shown in Fig. 20 (and Fig. 21 left) and has been removed for the image shown in Fig. 21, middle. Fig. 21 right shows the same image but with an offset added so that the mean value agrees with the mean value of the UPV image. As can be seen, bug correction and the added offset produce an image which agrees well with the other images including those that use a Combo-MF.

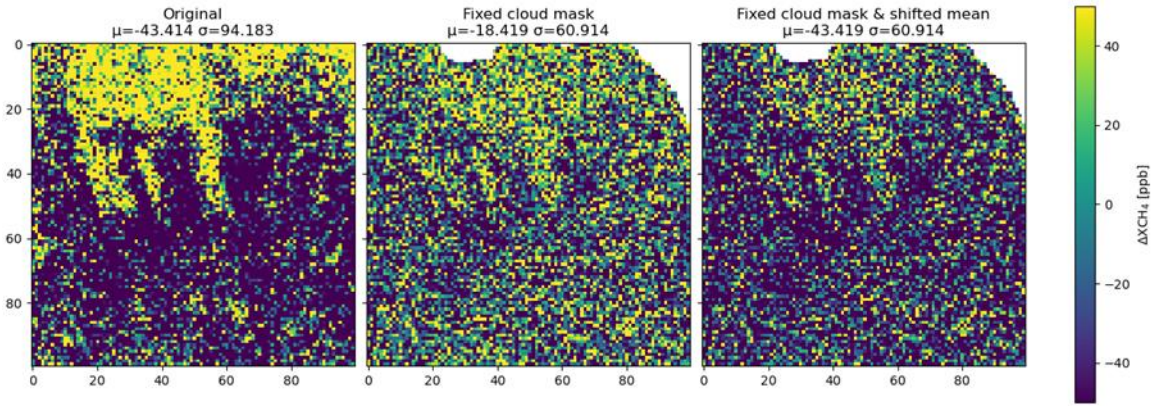


Figure 21. *Similar as Fig. 20 but limited to IUP-UB. Left: As IUP-UB in Fig. 20, which contains artifacts due to a bug related to cloud masking (affecting the covariance matrix as used for the MF method). Middle: Same as the left image, but with the cloud masking bug removed (white areas are due to cloud filtering). Right: As middle but with offset added (25 ppb) so that mean value is similar as the mean value of the other data sets (approx. -45 ppb).*

PRISMA – Madrid (Spain)

In Fig. 22, we find a similar results compared to the previous two scenes, therefore leading to similar conclusions. Moreover, we can see that there is a strong retrieval artifact in the middle of the subset that it is not even attenuated at background level, which is possibly due to a very high agreement with methane absorption features. On the other hand, we observe that IUP-UB uses a radiance quality filtering different from the rest of the groups leading to stripes of missing data. Using different pixels for the calculation of the statistic parameters required to carry out the matched filter is an additional source of discrepancy in the results.

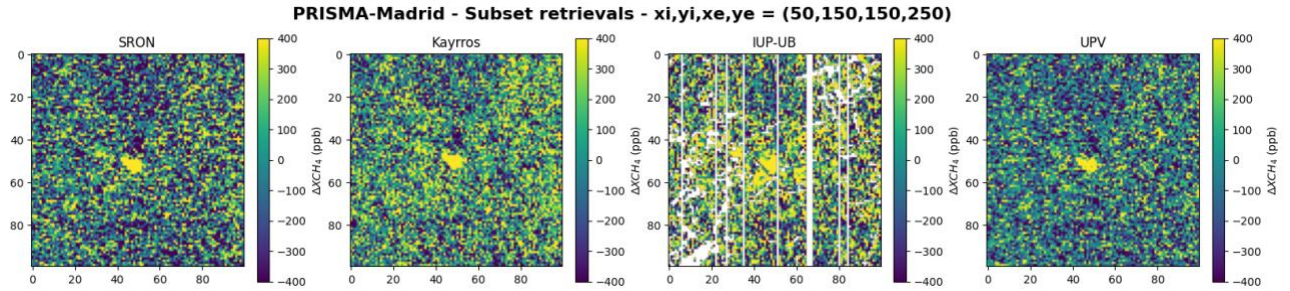


Figure 22. *Methane retrievals from the selected subset of the PRISMA acquisition capturing an area in Madrid (Spain). Starting from the left: SRON, Kayrros, IUP-UB, and UPV groups.*

Joint comparison: mean and standard deviation

The Anna Creek (Australia) and Shanxi (China) areas from these acquisitions are bright and homogeneous, while the Madrid (Spain) area is relatively dark and heterogeneous. Darker areas lead to a lower Signal-to-Noise-Ratio (SNR), which in turn results in poorer retrieval precision in the selected subset. Moreover, heterogeneity is also an important aspect for detection, since the matched filter assumes a completely homogeneous scene where spectra deviations in those bands that match methane's wavelengths only occur due to the existence of methane enhancements. Under a heterogeneous scene, less robust statistics are generated to carry out the matched filter, which will generally lead to a worse retrieval performance.

The use of one or another instrument also matters since their specific features have an important impact on the resulting retrievals. Even though EMIT has a similar spectral resolution to PRISMA and EnMAP, it has a substantially higher SNR that benefits retrieval precision (Thompson, 2024). On the other hand, EnMAP's and PRISMA's finer spatial resolution results in a higher level of retrieval detail that can also favor detection. Moreover, although PRISMA and ENMAP do show the same spatial resolution, the latter has proven to allow better retrieval performance (Roger, 2024).

We found 2 metrics to assess the performance of the detection retrievals, which are the mean and the standard deviation of the retrieval subsets. Since all subset pixels have no presence of methane, we can assume a Gaussian distribution with a mean value 0 and a standard deviation σ. We can measure the occurrence of retrieval artifacts with the mean value. If the value is different from 0, we can infer that there is a presence of artifacts that are making the mean to deviate from that value. The retrieval subsets are 100 x 100 pixels, so the effect of a retrieval artifact of just a few pixels can have a significant impact in this metric. On the other hand, σ is an indicator of retrieval precision and sets the level of background noise.

In Fig. 23, we find the mean (bars) and standard deviation (error bars) of all the retrieval subsets for all the acquisitions used in this activity for the 4 participating groups. Regarding the standard deviation, we can observe that the largest values occur for the Madrid (Spain) area, which is translated in a poorer retrieval precision. This is consistent with the characteristics of the area, since it is dark and heterogeneous. However, the results of Madrid for the EMIT instrument do not show standard deviation values as large as those from the PRISMA and EnMAP cases, which can be explained by the superior SNR that improves the retrieval precision. In fact, for every area, we can observe that those retrievals coming from EMIT data exhibit the best retrieval precision. In general terms, groups show consistency in the standard deviation levels with no important differences. The EnMAP-China and PRISMA-Madrid cases show the largest absolute mean values, which reflects the large number of pixels related to retrieval artifacts in the subset areas, as can be observed in Fig. 20 and 22, respectively. The mean values for all studied cases are negative for the SRON and UPV teams, which is a result of using the Combo-MF method due to its characteristic penalization. This leads to a generally lower fraction of positive enhancement values, which are the potential candidates for true methane enhancement pixels. As a result, it increases the ability to distinguish between true emissions and background.

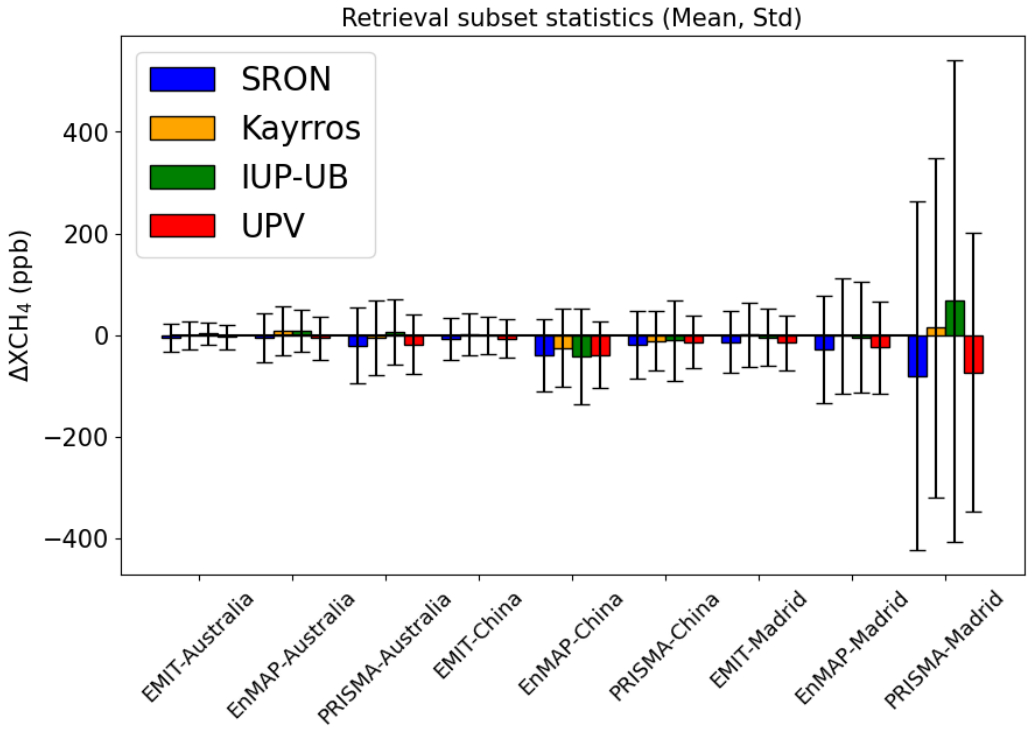


Figure 23. *Mean (bars) and standard deviation (error bars) values related to the SRON (blue), Kayrros (yellow), IUP-UB (green), and UPV (red) groups of all the retrieval subsets used for the assessment of methane concentration retrievals used for detection activity.*

Finally, we did not account for other factors that can also cause discrepancies such as the use of an iterative processing to account for the matched filter sparsity assumption or the calculation of the unit methane absorption spectra. Along with the spectral window selection, the radiance quality filtering, and the use of alternative matched filter flavors, these factors can explain most differences. However, our results indicate that the main differences came from the use of the particular matched filter flavor and the spectral window selection.

Moreover, Combo-MF reduces the occurrence of retrieval artifacts, which is helpful to filter them out of the plume mask. This is especially important for the IME quantification, since plume delineation is needed. Therefore, the selection of different matched filter flavors for plume delineation can indirectly have an impact on the flux estimates.

5.3. General methane emission quantification assessment

A list of methane point source emissions captured by EnMAP, PRISMA, and EMIT in several locations with various features have been quantified by all groups. Bar charts comparing the different group flux rate estimates (left Y-axis) for the different emissions are shown in Fig. 24, for EnMAP (left), PRISMA (center), and EMIT (right). Moreover, we can also compare the wind speed values used by each group in the quantification process, marked as pink stars (right Y-axis).

SRON, Kayrros, and IUP-UB use ERA5 to obtain the wind speed values, while UPV uses GEOS-FP. SRON uses GEOS-FP in cases where the plume direction is not properly aligned with ERA5 but is consistent with GEOS-FP. Beyond the use of one or another meteorological product to extract wind speed data, there are additional sources of variability. This is exemplified by the EnMAP-Turkmenistan quantification where the wind speed values vary for SRON, Kayrros, and UPV groups. These differences result from different wind processing. For instance, UPV uses a linear interpolation between 2 adjacent GEOS-FP U10 values (hourly resolution) to deduce the value related to 10 minutes before the acquisition time, while other groups apply another processing steps. As a result, we do not find agreement among group wind speed values. One of the most remarkable cases of wind speed variability can be found in the PRISMA-China case, where SRON, Kayrros, and IUP-UB exhibit a similar wind speed value around 3 m/s, while UPV's value is around 5 m/s. This difference might be a significant source of discrepancy in the group flux rate values, since they are directly correlated with wind speed. In fact, we find that UPV exhibits the highest Q in reference to the other groups for this case, which is probably due to this reason. However, to further investigate the main source of discrepancy for a specific emission case, we need to carry out a more comprehensive analysis that allows us to understand the impact of the used wind speed on the emission estimates (see Sect. 5.4).

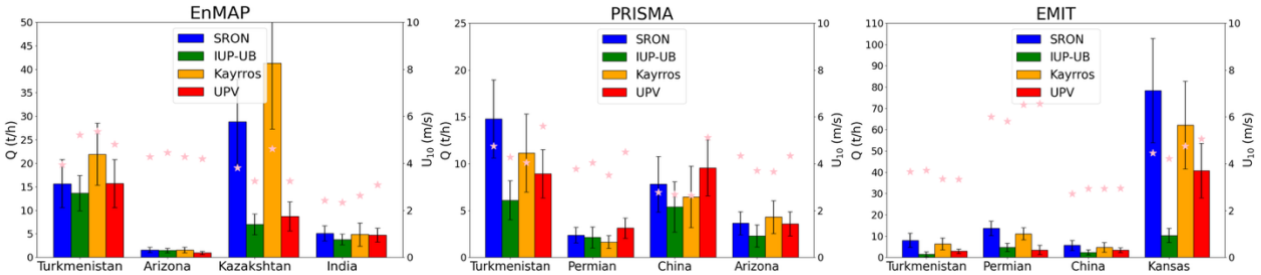


Figure 24. *Flux rate estimations (left y-axis, bars) and U10 values (pink stars, right y-axis) reported by SRON (blue), Kayrros (yellow), IUP-UB (green), and UPV (red) teams for several methane emission cases obtained using EnMAP (left), PRISMA (center), and EMIT (right) data.*

Regarding the Q values reported by the teams, we also observe some variability, which can be expected due to the large emission rate uncertainty. Along with the wind speed value variability, the use of different methane enhancement maps and quantification steps are leading to discrepancies. In some cases, we find remarkable differences, such as in the one from EMIT-Kansas, where differences in Q values can be as large as one group having a

value that is 8 times higher than another. In addition, we observe that in 9 out of the 12 emissions IUP-UB exhibits the lowest Q value. While the other groups use the IME quantification method, IUP-UB is the only one that uses the CSF method. It needs to be investigated why the IUP-UB CSF results are lower compared to the other results. Nevertheless, except for a few cases, we generally find that group estimates agree within the uncertainty bars.

5.4. Detailed methane emission quantification assessment

To have a better understanding of the emission rate discrepancy among groups, we will assess the retrievals used for quantification and the influence of the different factors in the group algorithm pipelines. We will use 3 plumes from the previous section, which have been selected to include all the instruments and preserve a certain degree of variability in emission and scene features. Next, we list the selected emissions along to some of their related features:

- EMIT – Kansas: large disagreement case, strong and long plume, dark and heterogeneous scene.
- EnMAP – Arizona: small plume, bright and heterogeneous scene
- PRISMA – Arizona: atypical plume shape, dark and heterogeneous scenes. Even though both PRISMA and EnMAP plumes are captured in Arizona, the specific locations differ.

Some groups developed new algorithm versions that are different from the ones used in previous activities. These groups showed higher confidence in the new versions, which in turn makes them more representative of the state-of-the-art methods. Therefore, we will use the new versions for this section. Note that the description of the different group algorithm versions can be found in Sect. 3.

5.4.1. Methane retrievals used for quantification

Groups provided unorthorectified retrievals for the 3 selected plumes, specifically those retrievals used for quantification. The goal of this activity is to carry out a per-pixel comparison to assess differences between retrievals, which ultimately has an impact on emission rate estimates.

Next, we will show subsets of the methane retrieval maps from each group capturing the plume area for the 3 different cases. Moreover, some scatter plots along with their related linear fits will used to compare the retrieval values between groups, focusing on (high) plume values. Retrieval artifacts can also lead to high scores, but we selected the retrieval subsets to minimize their appearance. We will analyze trends and relationships based on these figures.

EnMAP – Arizona

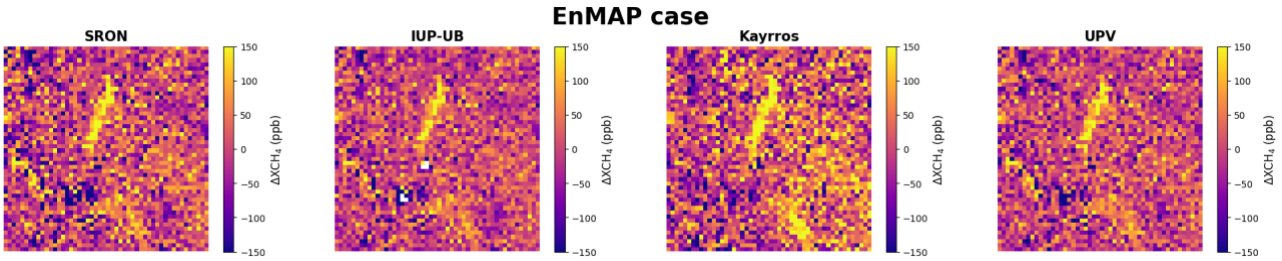


***Figure 25**. Subsets of methane concentration enhancement maps from each group capturing the plume characterized with EnMAP data.*

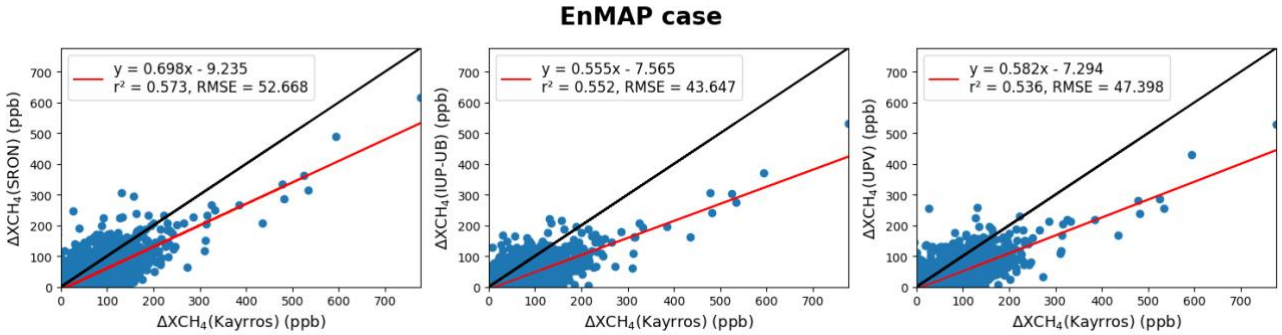


***Figure 26**. Scatter plots comparing the retrieval values from Kayrros to those from SRON (left), IUP-UB (center), and UPV (right) for the plume captured with EnMAP data.*

In Fig. 25, we can see considerably larger values for the Kayrros retrieval in reference to the rest of the groups. This is also observed in the linear fits from Fig. 26, which indicate that the values from UPV and IUP-UB are approximately half those from Kayrros. SRON values are in between. Moreover, UPV and IUP-UB values seem to show good agreement.

PRISMA – Arizona

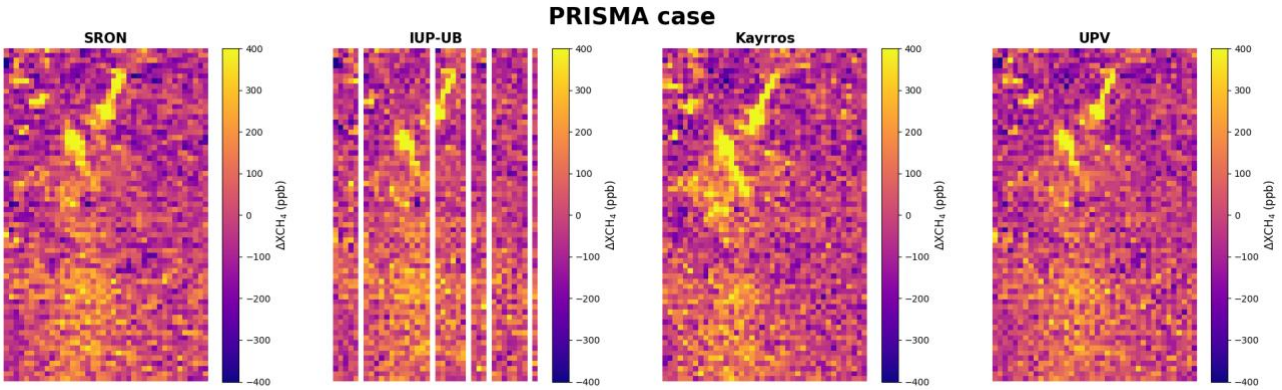


***Figure 27.** Subsets of methane concentration enhancement maps from each group capturing the plume characterized with PRISMA data.*

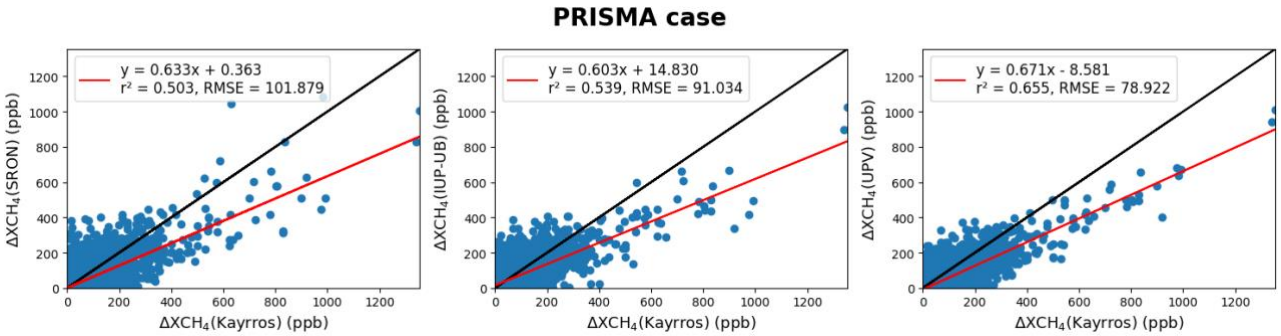


***Figure 28**. Scatter plots comparing the retrieval values from Kayrros to those from SRON (left), IUP-UB (center), and UPV (right) for the plume captured with PRISMA data.*

In Fig. 27 and Fig. 28, we again see larger values for the Kayrros retrieval and good agreement between IUP-UB and UPV. However, the underestimation of IUP-UB and UPV in reference to Kayrros is not as pronounced as in the EnMAP case. Moreover, SRON values seem to better agree with those of IUP-UB and UPV.

EMIT – Kansas

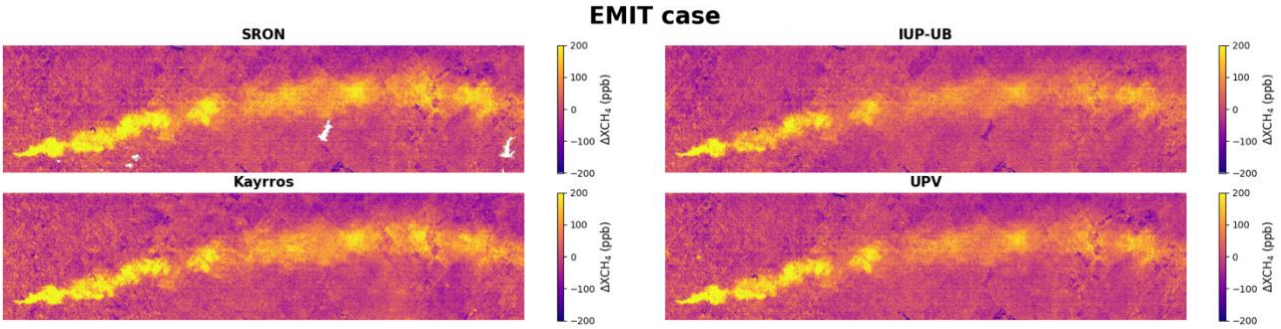


***Figure 29**. Subsets of methane concentration enhancement maps from each group capturing the plume characterized with EMIT data.*

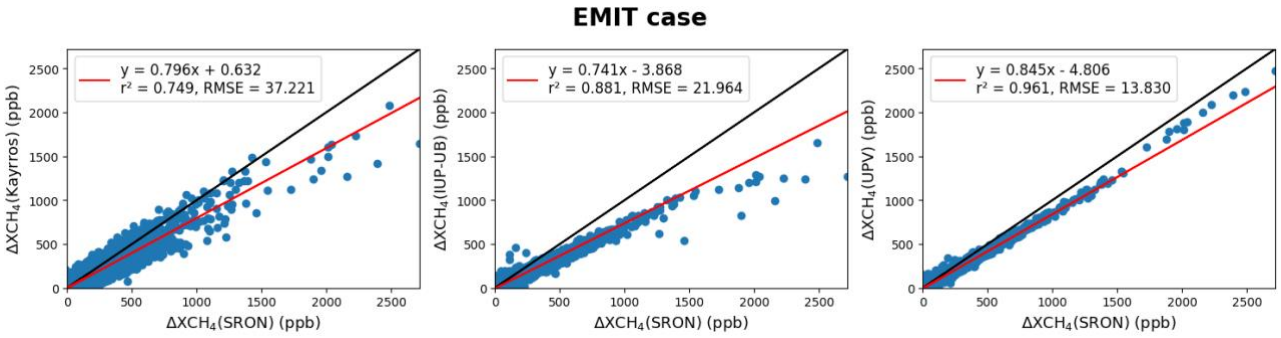


***Figure 30**. Scatter plots comparing the retrieval values from SRON to those from Kayrros (left), IUP-UB (center), and UPV (right) for the plume captured with EMIT data.*

In Fig. 29, we see that now SRON reports higher retrieval plume values compared to the rest, instead of Kayrros. This can be better observed in Fig. 30, since the linear fits show higher values for SRON in all cases. After a careful assessment, we also observed that IUP-UB and UPV values approximately align in the plume tail, while higher values are found near the source for UPV.

Overall, we find the following points after considering the 3 different plumes:

- For EnMAP and PRISMA cases, Kayrros plume pixel values are consistently higher than the ones from the rest of the groups. On the other hand, for the EMIT case, SRON plume pixel values are higher than those from the rest of the groups. The more obvious difference is that the EMIT plume is significantly stronger.
- For all cases, IUP-UB and UPV plume pixel values are similar, except near the source for the EMIT case, where enhancement values are larger.
- For the EnMAP and EMIT cases, SRON plume pixel values are higher than those from UPV and IUP-UB (yet lower than those from Kayrros). However, values are approximately similar in the PRISMA case.

At this point differences in retrieval plume values can arise from several factors such as the selection of the spectral window used in the retrieval method, the matched filter flavor, or differences in the methane unit absorption spectrum (K). Next, we apply specific analysis to some cases to understand why we find discrepancies.

For the EMIT case, SRON and UPV use a similar spectral window and the same matched filter flavor, i.e. the lognormal matched filter. Then, to better understand differences between these two retrievals, we will analyze the impact of the selection of K. In absolute values, this spectrum illustrates the fraction of radiance that is absorbed by methane per unit of concentration. Therefore, considering the same atmospheric vertical column, a K showing deeper absorption features would lead to lower retrieval values than a K with less pronounced features. In Fig. 31, we can see the K spectra used by both groups. Note that UPV uses two different K spectra depending on the retrieval enhancement values (see Sect. 3.3). We can observe that UPV spectra exhibit more pronounced absorption features than the SRON spectrum. Therefore, UPV will result in lower retrieval plume values, which aligns with what we observed in Fig. 30. UPV has run again the retrieval for the EMIT case following their original algorithm, but this time considering the SRON K spectrum. In Fig. 32, we observe that now there is a good agreement between SRON and UPV with a slope close to 1. Therefore, we can conclude that the main source of discrepancy between these groups is the selection of the K spectrum. Hence the larger values from SRON compared to the other groups in the EMIT case are likely mainly due to the selection of the K spectrum.

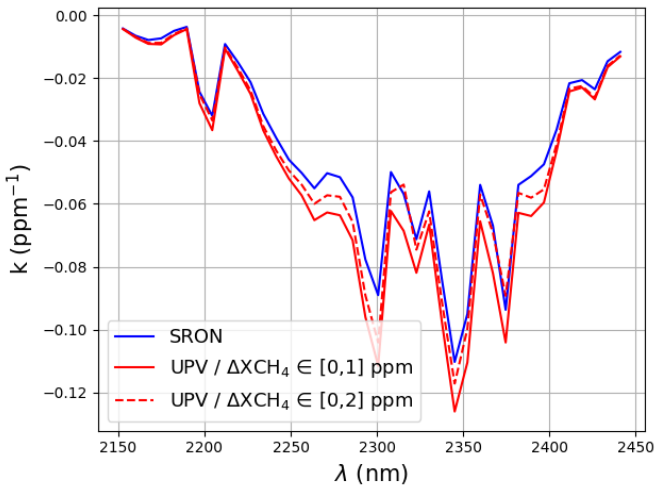


***Figure 31**. K spectra used by SRON (blue) and UPV (red) for the EMIT case. UPV considers a K derived from ΔXCH₄ values between 0 and 1 ppm (solid line) and between 0 and 2 ppm (dashed line).*

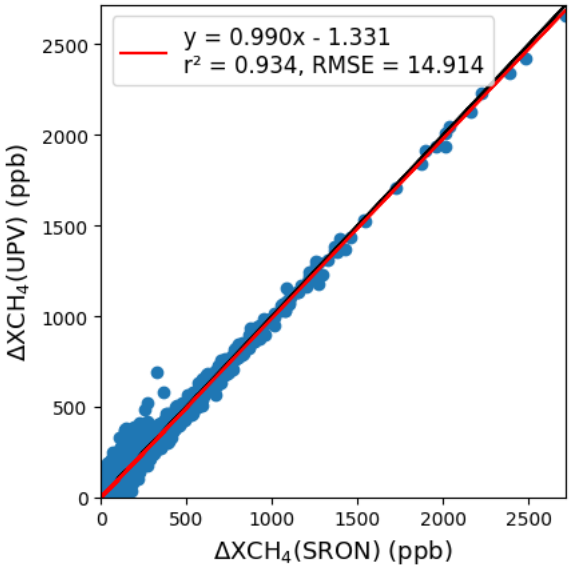


***Figure 32**. Scatter plot comparing the retrieval values from SRON and UPV using the SRON K spectrum for the plume captured with EMIT data.*

There are several aspects that could cause differences between the spectra such as the selection of the ΔXCH_4 range used to the deduction of K, the selection of the atmosphere type, or the consideration of applying a correction to the convolved ISRF spectra (Gorroño et al., 2023). We found the following differences between both groups:

- SRON considered a midlatitude winter atmosphere, while UPV used a US Standard one.
- UPV convolved the radiometric spectra from their LUT to the ISRF and then applied a correction. However, this correction is not applied by SRON.
- SRON used a ΔXCH_4 value range from 0 to 6.4 ppm, while UPV used 2 different ranges. The first one is from 0 to 1 ppm, and the second one is from 0 to 2 ppm.

After setting up the same type of atmosphere for both groups, we found that there are no remarkable differences. However, we did find a lower discrepancy in retrieval values when combining both the same ΔXCH_4 value range and setting up both groups without correction. In Fig. 33, we observe closer K spectra values between the 2 groups (left panel) and a linear fit closer to the 1:1 relationship (right panel), where differences were reduced from 15% to 6%.

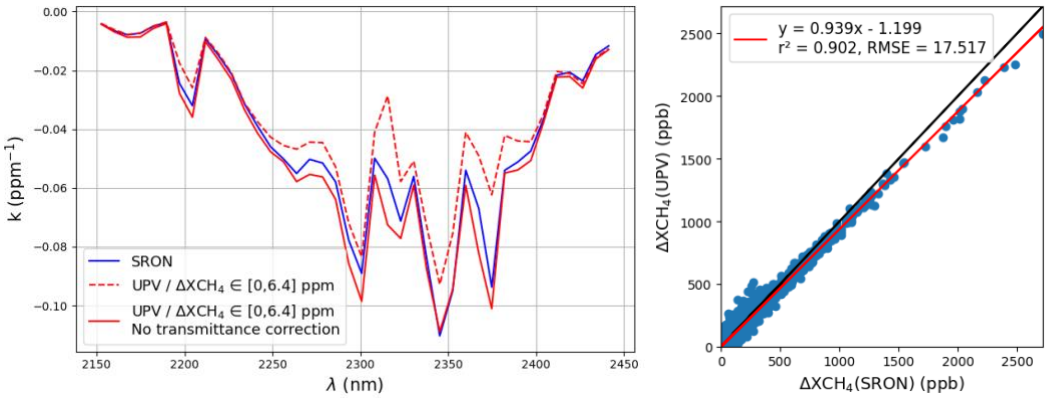


***Figure 33**. K spectra (left panel) used by SRON (blue) and UPV (red), where UPV considers a K derived following the same ΔXCH₄ sampling as SRON and applying a correction to the convolved ISRF spectra (dashed line) and not applying a correction (solid line). In the right panel, a scatter plot comparing the retrieval values from SRON and UPV, where UPV uses the latter described K spectrum.*

At this point, remaining differences might originate from deeper differences in the used Radiative Transfer Models and how the group LUTs are built. Thus, we observed that the selection of the K spectrum can be an important source of differences. In future work, differences in how the K spectrum is obtained should be carefully analyzed.

Regarding the EnMAP case, we observed systematic higher retrieval plume values for Kayrros in reference to the rest of the groups (see Fig. 26). Moreover, we also observed that SRON values were also higher than those from UPV. We compare the corresponding K spectra (see Fig. 34) and observe that Kayrros' K spectrum exhibits less pronounced absorption features than SRON, and the latter shows less pronounced features than UPV. This perfectly matches with the difference in retrieval plume levels, which again shows the importance of selecting the K spectra to obtain methane retrievals. Kayrros higher values for the PRISMA case (see Fig. 28) are likely to also be a result of the selection of K:

- Kayrros uses the 1976 US standard atmosphere from LOWTRAN 7 atmospheric transmittance and background radiance model (Kneizys et al., 1988).
- No scene-based spectral recalibration is applied.

This could explain the difference observed when comparing Kayrros's results to those from the other groups.

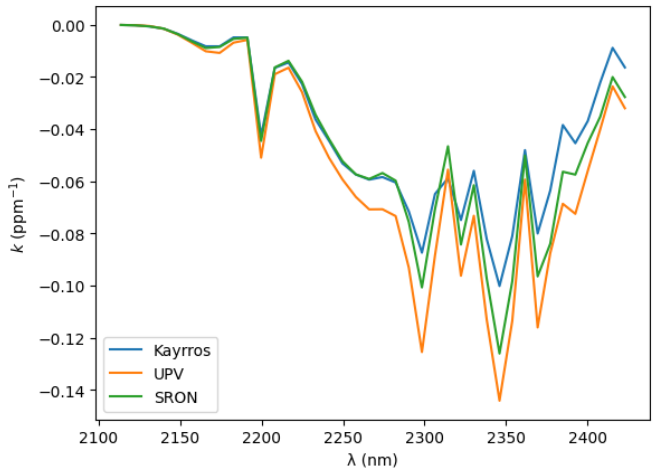


***Figure 34**. K spectra used by Kayrros (blue), UPV (orange), and SRON (green) for the EnMAP case.*

5.4.2. Comprehensive quantification assessment

Groups provided the emission rate estimates for the 3 selected plumes considering 4 different cases:

- No common products.
- Common wind speed (set to U_{10} = 3 m/s).
- Common retrieval (different from every original group retrieval), where the retrievals correspond to the methane concentration enhancement maps.
- Common wind speed and retrieval.

In this manner, we analyze the main factors that make group estimations differ. If group estimations do not agree after fixing the wind speed and retrieval, then we can relate differences to the quantification process.

EnMAP – Arizona

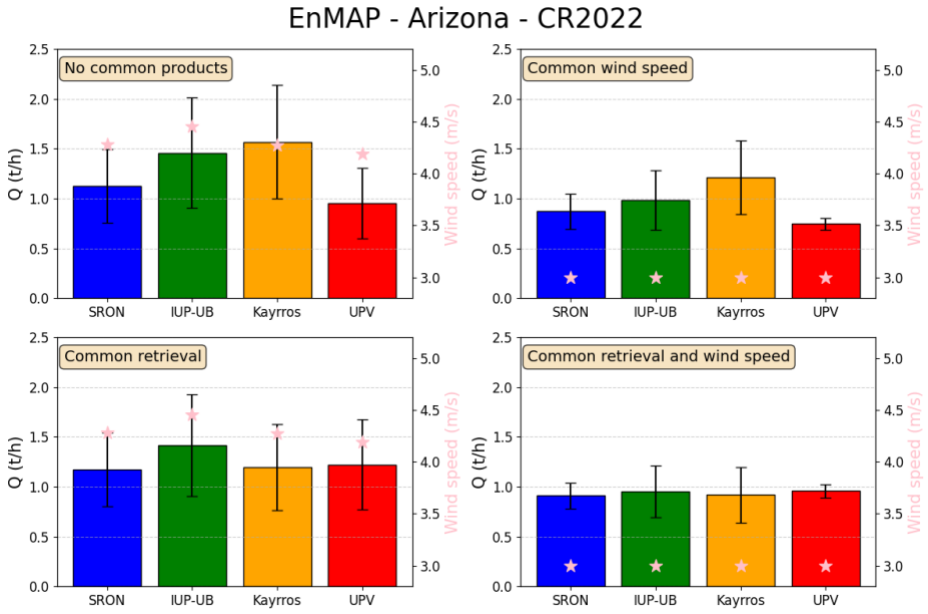


***Figure 35**. Emission rate estimations (bar charts) with their related errors (error bars) for the 4 different quantification cases along the used wind speed values (pink stars) for a plume located in Arizona and characterized using EnMAP data.*

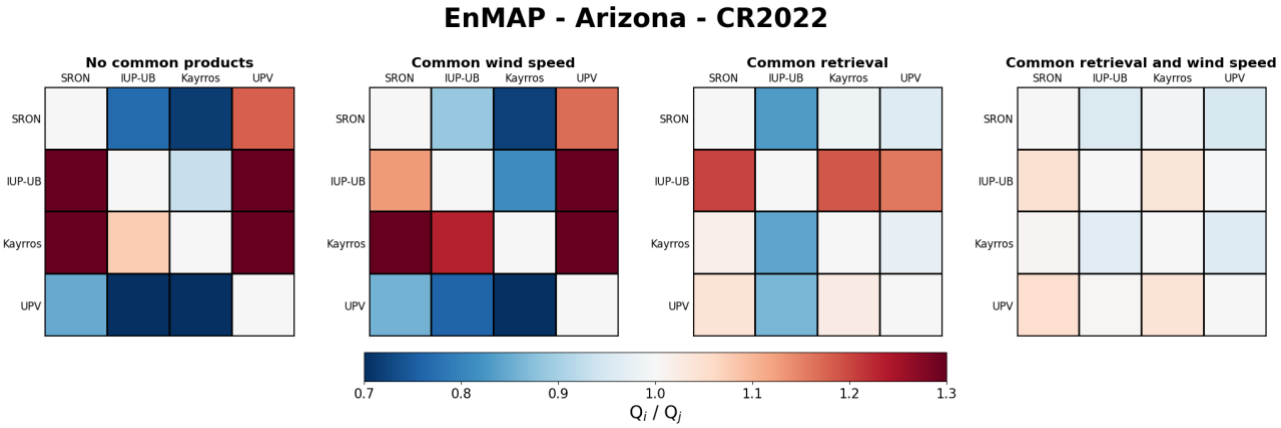


***Figure 36**. Matrix of the emission rate ratio (group from row divided by group from column) for the 4 different cases for a plume located in Arizona and characterized using EnMAP data.*

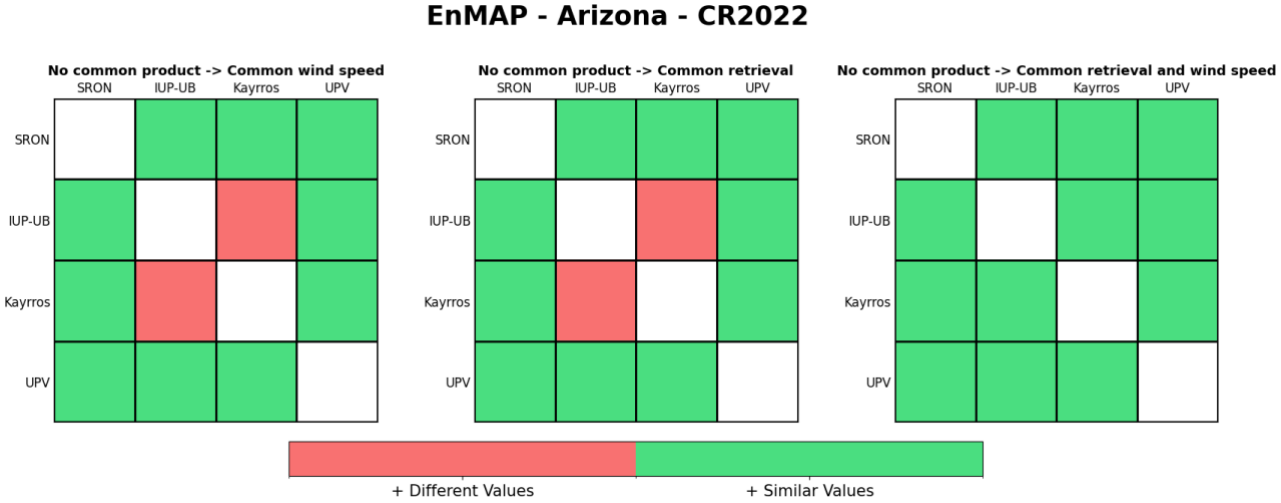


***Figure 37**. Matrix representing the increase or reduction of discrepancy among group estimations in the transition between the 'No common product' case to the 'Common wind speed' (left), 'Common retrieval' (center), and 'Common retrieval and wind speed' (right) cases for a plume located in Arizona and characterized using EnMAP data.*

No common product

In the 'no common products', there are discrepancies among groups, although all values match within the error bars. At this point, differences can arise from the retrieval, the wind speed, and the quantification method. For instance, within the IME quantification method, plume masking criteria and the U_{eff} calibration play an important role in reference to discrepancies.

Common wind speed

If we only fix the wind speed, values are reduced since the fixed wind speed (U_{10} = 3 m/s) is lower than every group's wind speed value. The flux rate depends on the U_{eff} for the CSF and IME methods. U_{eff} decreases for lower U_{10} values (Varon, 2018), which makes the observed reduction consistent. Moreover, we can observe some of the following changes:

- The relative difference between IUP-UB and Kayrros increases.
- IUP-UB has a U_{10} ~ 4.5 m/s, while Kayrros has a U_{10} ~ 4.3 m/s.
- After normalizing the wind speed, we find a relatively lower Q for IUP-UB.

Visually, it is difficult to tell the evolution of the relative differences among the rest of the groups if we only look at Fig. 35. Wind speed values used by groups range between 4.1 m/s and 4.5 m/s, which is a relatively low variability regarding the typical large uncertainty since these values are retrieved from low resolution products. Therefore, big differences are unlikely to be found after normalizing the wind speed. For instance, regarding those groups that use the IME quantification, we consider the U_{eff} implemented by UPV ($U_{eff} = 0.34 \cdot U_{10} + 0.44$), who used a U_{10} ~ 4.2 m/s. On the other hand, SRON used a U_{10} ~ 4.3 m/s and a similar U_{eff} ($U_{eff} = 0.38 \cdot U_{10} + 0.43$). Then, the factor of reduction of UPV Q after fixing the wind speed (U_{10} = 3 m/s) is 0.78, which is very similar to the 0.76 from SRON. Kayrros also uses an IME quantification and a similar effective wind speed calibration, which will lead to similar changes. Nevertheless, the more subtle differences can be better seen in the other 2 figures, where we show the emission rate ratio among groups (Fig. 36) and whether there is an improvement in reference to the 'no common product' case (Fig. 37).

Common retrieval

If we only fix the retrieval, we see that UPV, Kayrros, and SRON show strong agreement. While UPV Q increased, Kayrros Q decreased. This can be linked to different plume pixel related enhancement levels, which is consistent with the retrieval analysis that was previously carried out. Considering that these 3 groups use the IME quantification, exhibit a similar U_{eff}, and use similar wind speed values, we can infer that plume masks from the groups are consistent. However, even though this happens for this particular plume, we cannot always expect plume mask consistency (see Sect. 5.1).

Common retrieval and wind speed

We generally do not find remarkable differences in reference to the 'common retrieval' case for most cases, since the group wind speed values considered are similar. However, when fixing both wind speed and retrieval, the IUP-UB quantification better matches the IME groups in reference to only fixing the retrieval. Then, we find consistency among all the groups' flux rates even though different quantification strategies are involved. We summarize the main points from this plume case as:

- Already without common products the four groups report consistent emission rates.
- SRON, Kayrros, and UPV: main differences are due to the retrieval.
- IUP-UB finds the highest consistency with SRON, Kayrros, and UPV after fixing retrieval and wind speed (joint contribution).

PRISMA – Arizona

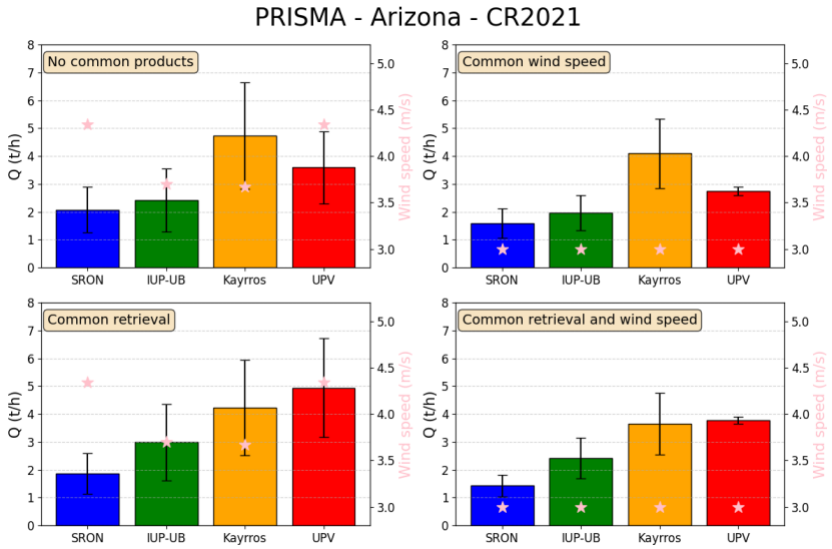


***Figure 38**. Emission rate estimations (bar charts) with their related errors (error bars) for the 4 different quantification cases along the used wind speed values (pink stars) for a plume located in Arizona and characterized using PRISMA data.*

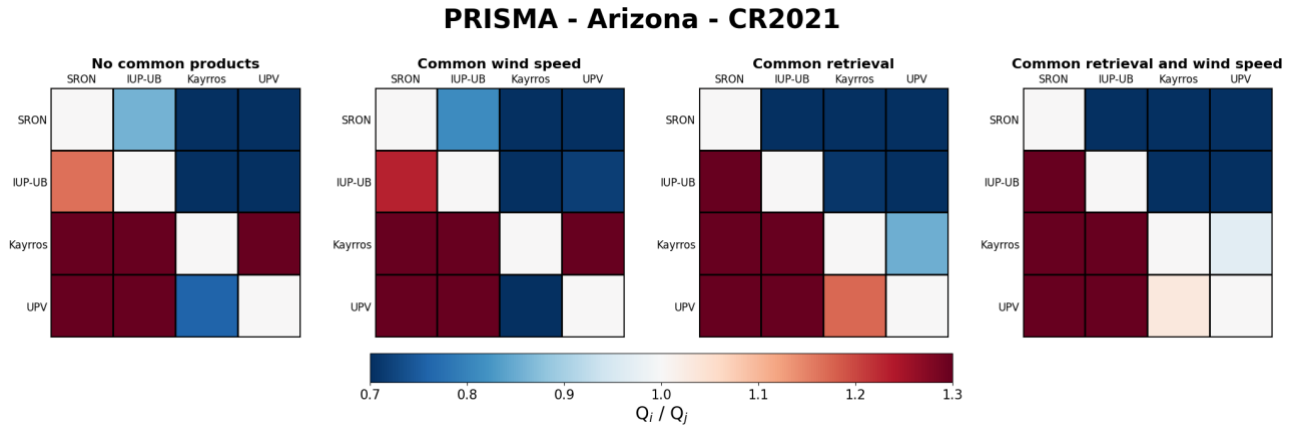


***Figure 39**. Matrix of the emission rate ratio (group from row divided by group from column) for the 4 different cases for a plume located in Arizona and characterized using PRISMA data.*

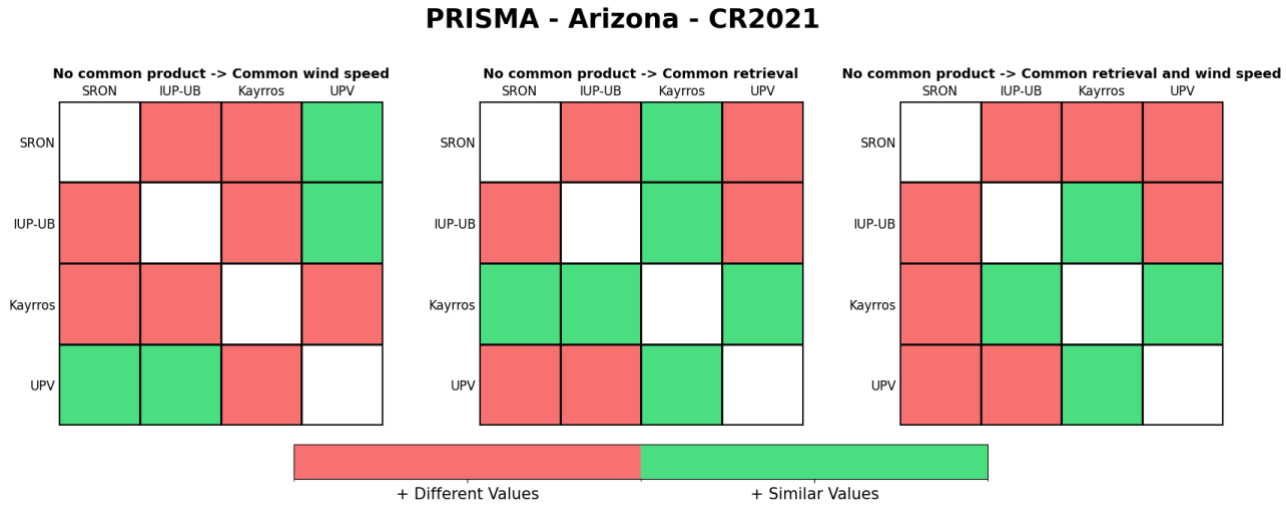


***Figure 40**. Matrix representing the increase or reduction of discrepancy among group estimations in the transition between the 'No common product' case to the 'Common wind speed' (left), 'Common retrieval' (center), and 'Common retrieval and wind speed' (right) cases for a plume located in Arizona and characterized using PRISMA data.*

No common product
Similarl to the EnMAP plume, all group Q estimations match within the error bars.

Common wind speed
In this plume case, IUP-UB and Kayrros show similar wind speed (~ 3.7 m/s). This happens similarly to SRON and UPV (~ 4.4 m/s). Therefore, after normalizing wind speed values, we expect to see the most remarkable changes among groups with different wind speeds. Moreover, maximum U_{10} differences are ~ 0.7 m/s, which is larger than the variability from the EnMAP plume case. Nevertheless, we do not find significant changes by only fixing the wind speed.

Common retrieval
Group pairs IUP-UB – Kayrros and SRON – UPV show similar wind speeds between each other. Therefore, if we fix the retrieval for these pairs, the remaining discrepancies can only be explained by the quantification. We first compare those groups with similar wind speeds:

- SRON – UPV: we observed that the discrepancy grows larger, which is probably because of differences in plume masking and the U_{eff} calibration. U_{eff} calibration for point sources using PRISMA data are very similar (see Sect. 3). Therefore, most differences must arise from plume masking.
- Kayrros – IUP-UB: the discrepancy is reduced considerably. It seems that the retrieval was one of the factors that made these two groups differ. Remaining differences can be attributed to the quantification method.

As for the remaining group combinations, we can highlight the following:

- Regarding IUP-UB – UPV and IUP-UB – SRON, we observe higher discrepancy in reference to 'common product' and 'common wind speed' cases. Therefore, differences cannot be explained by only fixing the wind speed or the retrieval.

- There is a slight improvement between Kayrros and UPV. Kayrros Q is higher than UPV's before fixing the retrieval, but this changes after fixing the retrieval. This is consistent with the retrieval analysis previously carried out, where we generally observed considerably higher plume pixel enhancements with the Kayrros algorithm.
- Discrepancy is reduced between SRON and Kayrros after fixing the retrieval.

Common retrievals and wind speed values
At this point, differences can only arise from the quantification process. We can observe some of the following points:

- Approximate agreement among Kayrros, and UPV. Non-negligible discrepancy was found after individually fixing wind speed and retrieval. However, a very good agreement was obtained after fixing both.
- Lower values reported by SRON in reference to the rest of the groups. After fixing wind speed and retrieval, the discrepancy generally grew larger. Therefore, a large fraction of the discrepancy comes from the quantification process. This likely occurs due to the atypical plume shape from the PRISMA case. Different plume masking criteria might have led to these strong biases.

The more restrictive mask applied by SRON was chosen to not have a negative impact on the estimated emission rate from a nearby albedo feature. To better understand the influence of plume masking in the flux rate discrepancy, SRON repeated the quantification using a different masking. The original SRON estimate, which is shown in Fig. 38, was obtained using a plume delineation from mask 1, while their second estimate was obtained using mask 2. Next, UPV repeated the quantification process using a similar plume delineation similar to that of SRON mask 2. Similarly, we will denominate the plume mask from the original UPV estimation as mask 1, and the second one as mask 2. In the left panel from Fig. 41, we can see both mask 1 from SRON (blue) and UPV (red), which are significantly different. In the center panel, both mask 2 are illustrated and they look more alike. After fixing the retrieval and the wind speed, we again show the groups' emission estimates (right panel) but also including the results from the UPV and SRON that were obtained using the plume delineations from their corresponding mask 2. We observed that, under a similar plume delineation, SRON and UPV exhibit similar Q values around 2 t/h. These results explain one of the main sources of discrepancy for the emission rate estimate for this particular plume and, in turn, highlight plume delineation as a key factor that significantly influences the quantification process when using the IME method.

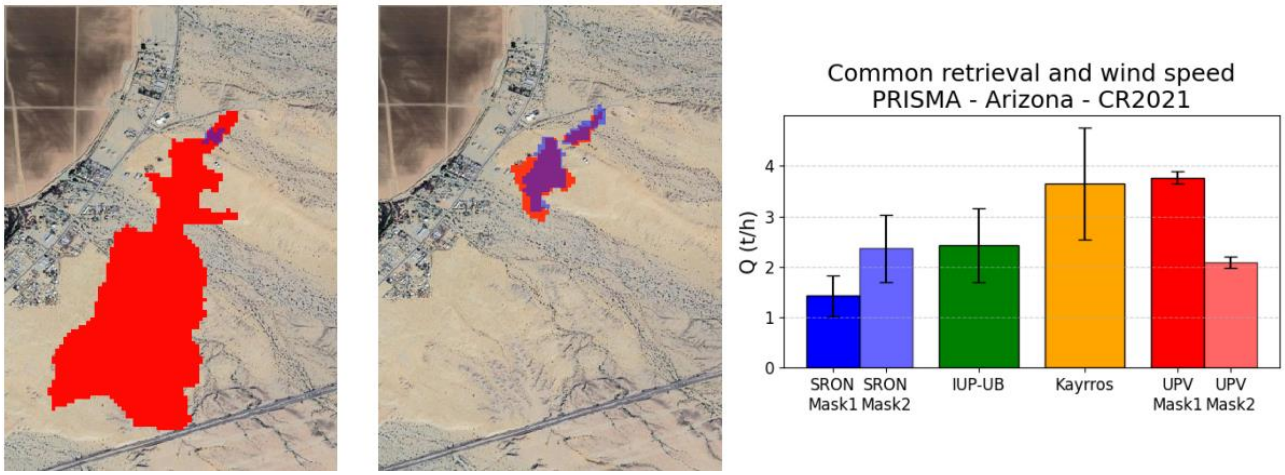


***Figure 41**. Mask 1 (left panel) and Mask 2 (center panel) of SRON (blue) and UPV (red) for the PRISMA case. On the right panel, the quantification of all the participating groups for this case, including both estimates from UPV and SRON using the different plume delineations.*

EMIT – Kansas

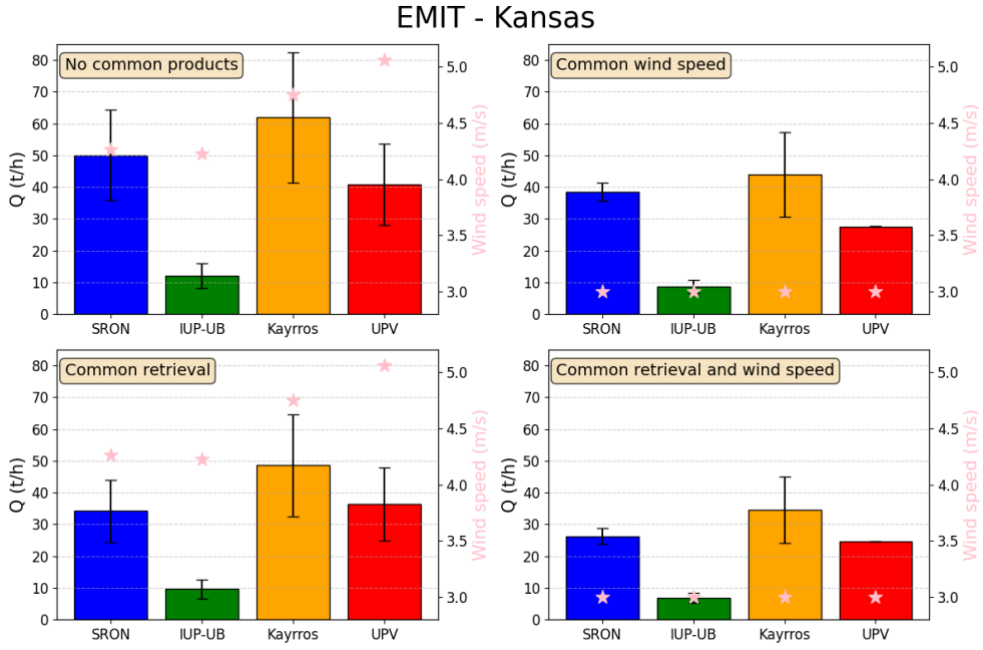


***Figure 42**. Emission rate estimations (bar charts) with their related errors (error bars) for the 4 different quantification cases along the used wind speed values (pink stars) for a plume located in Kansas and characterized using EMIT data.*

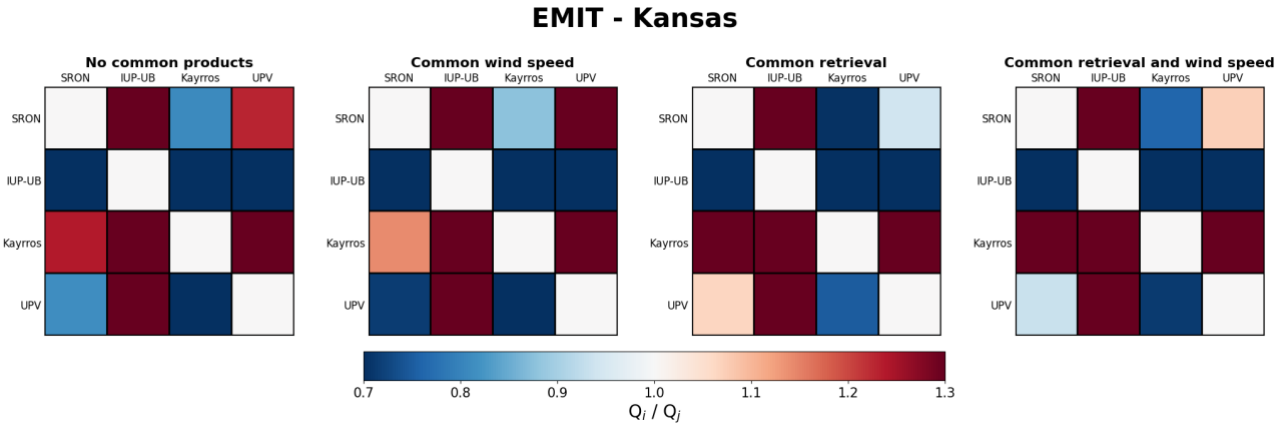


***Figure 43**. Matrix of the emission rate ratio (group from row divided by group from column) for the 4 different cases for a plume located in Kansas and characterized using EMIT data.*

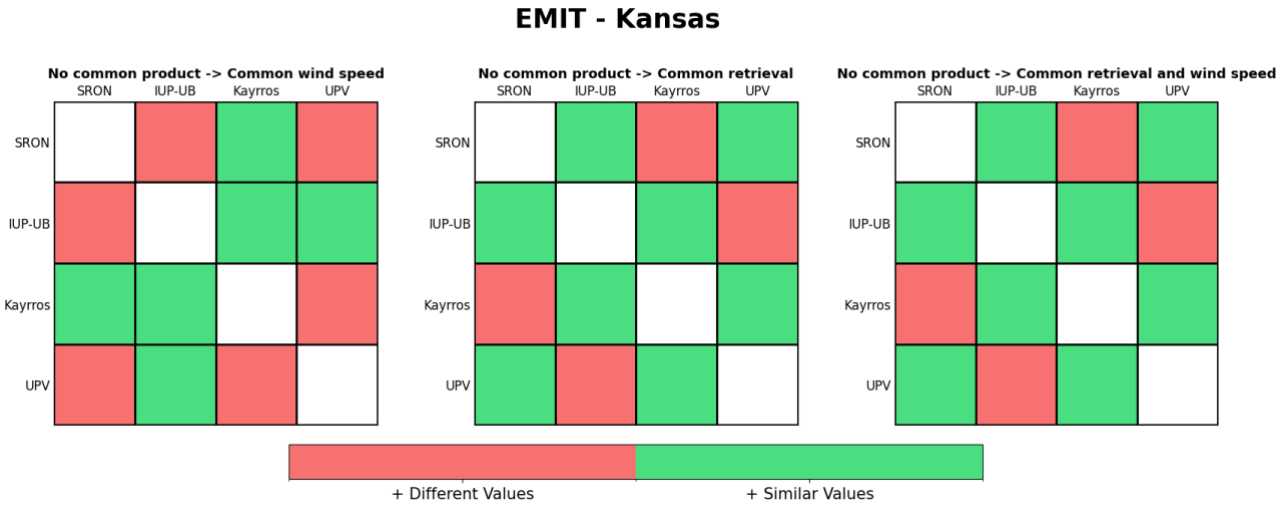


***Figure 44**. Matrix representing the increase or reduction of discrepancy among group estimations in the transition between the 'No common product' case to the 'Common wind speed' (left), 'Common retrieval' (center), and 'Common retrieval and wind speed' (right) cases for a plume located in Kansas and characterized using EMIT data.*

No common product

All values match within the error bars, except for the IUP-UB case, which shows large underestimation in reference to the rest of the groups. To investigate this discrepancy, SRON applies their CSF method to show the emission rate's sensitivity to the plume length. The mean emission rate derived from the CSF method agrees with their IME result. However, if they limit the plume length to 500 m (~ 8 lines), which is applied in IUP-UB's CSF method, the emission rate would be 28 t/h. This could explain why IUP-UB's result is much lower than other groups' IME results.

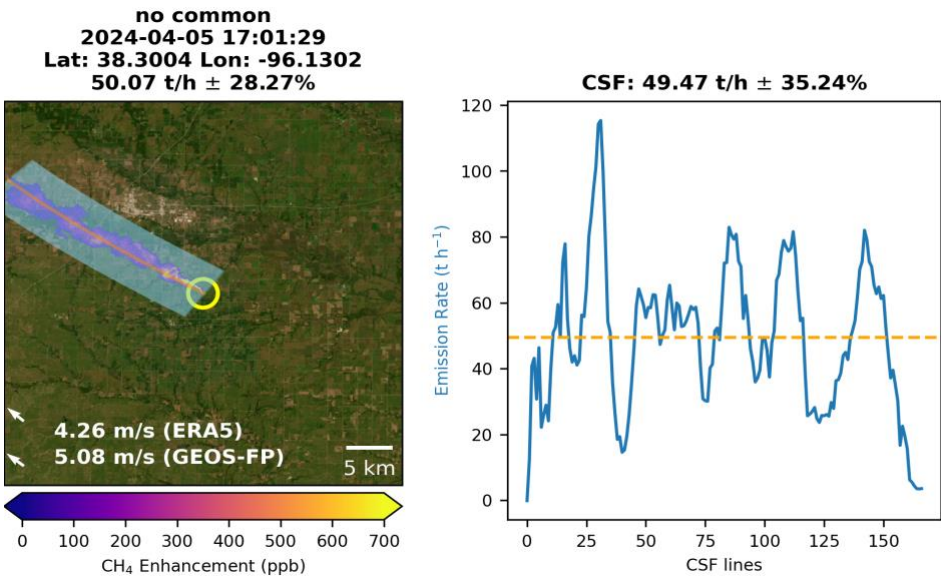


Figure 45. *The SRON IME (left) and CSF (right) results for the EMIT case. The CSF crosslines (blue lines) are shown on the left, while the emission rate per line is shown on the right. The titles are estimates and uncertainties for each method.*

On the other hand, IUP-UB has carried out additional analysis to investigate possible reasons for underestimation of the IUP-UB CSF-based emission estimates for the EMIT/Kansas case as shown in the upper left panel of Fig. 42. To investigate this, several parameters of IUP-UB's CSF algorithm have been varied including the size of the spatial box used for CSF analysis (determining the part of the plume that is analyzed) and the background correction. For these variations, also quite extreme cases have been included, such as enhancing the box size from 0.5 km (current baseline, see Fig. 46) to 24 km (Fig. 47) and setting all background values to zero. For most cases, the resulting fluxes were similar as the reported value of 12 ± 4 t/h as obtained with the baseline setting. Differences larger than approximately a factor of 2 have not been found. The maximum emission was 25 ± 10 t/h, which is still significantly lower compared to the results of the other groups, which are in the range 40 - 80 t/h. The assessment also included the application of the IUP-UB CSF method to the Level 2 (delta XCH4) image from UPV instead of the IUP Level 2 image. Also, this resulted in only small changes of the derived emissions. From these assessments it is concluded that the results obtained with the IUP-UB CSF method are quite stable with respect to algorithm related settings and Level 2 input data.

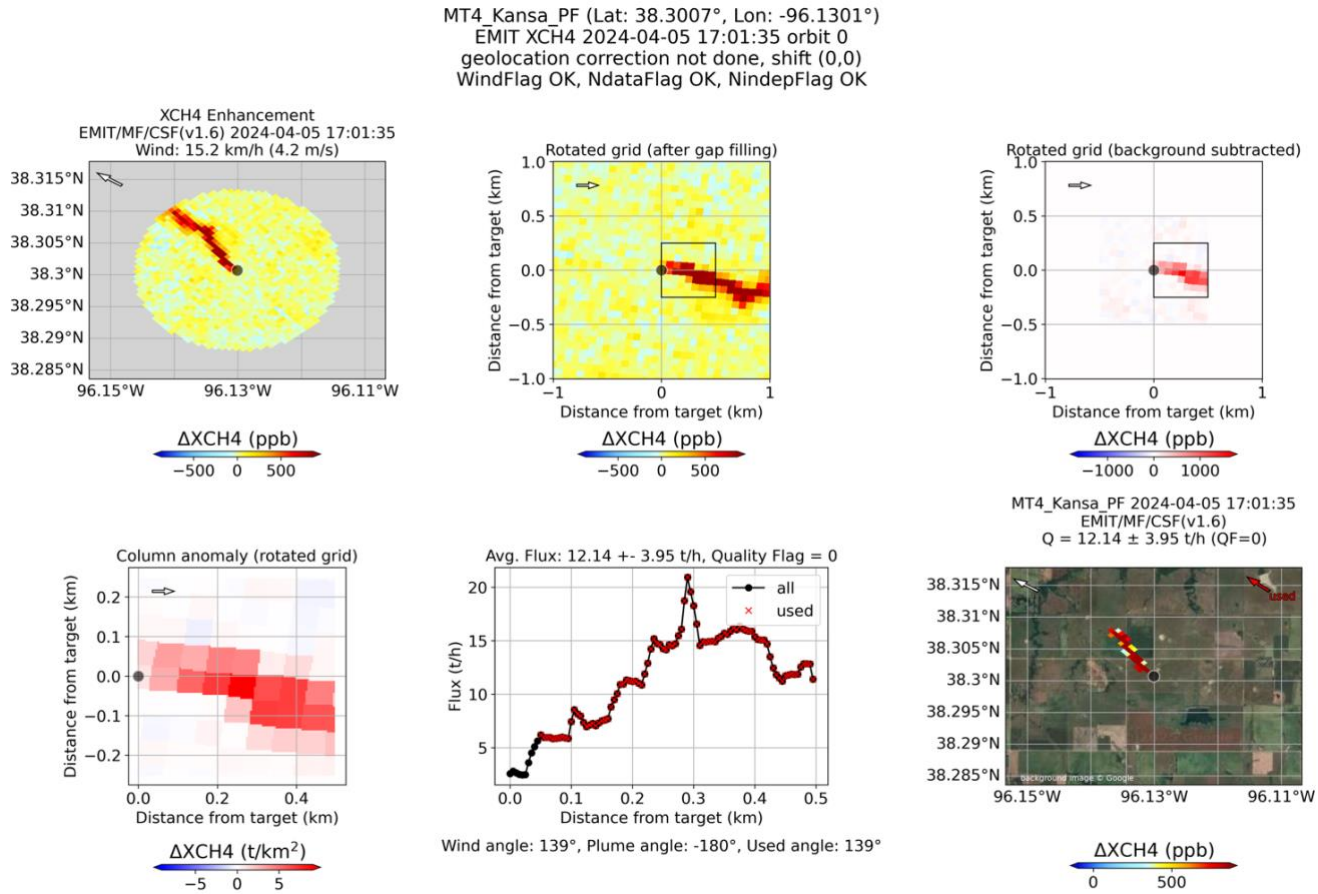


Figure 46. *The IUP-UB results for the EMIT case using default settings, e.g., box size 0.5 km. The estimated emission is 12 +/- 4 t/h.*

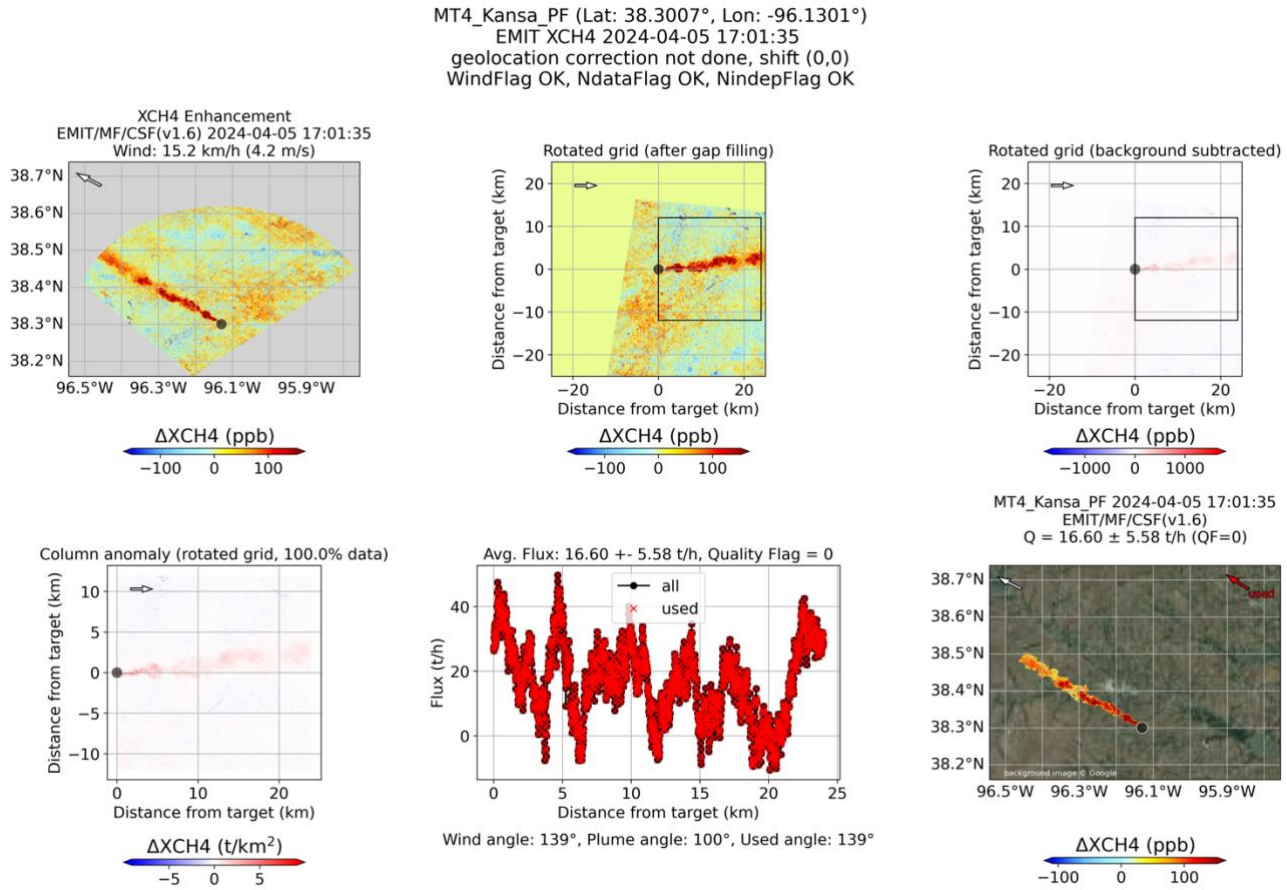


***Figure 47**. As Fig. 46 but using a much larger box size (24 km instead of 0.5 km), i.e., a larger plume. Here the estimated emission is 17 +/- 6 t/h.*

After the SRON and IUP-UB analysis regarding the flux estimation variability in the CSF quantification method, we can conclude that the selection of the involved parameters can lead to high variability in the results. After the modification of these parameters, IUP-UB obtained a value of 25 t/h, which is twice their original estimate. In contrast, limiting the SRON-CSF approach to the first 500m of the plume led to an estimate of 28 t/h, which is nearly half of their value applying the CSF to the entire plume. Therefore, there is a significant variability in the results depending on these parameters and, under a specific selection, both group estimates exhibit higher consistency.

Common wind speed

In this plume case, SRON and IUP-UB show similar wind speed (~ 4.2 m/s). However, Kayrros and UPV show U_{10} values around 4.7 m/s and 5 m/s, respectively. Therefore, maximum differences among groups regarding wind speed are ~ 0.8 m/s, which is approximately similar to the variability from the PRISMA plume case. We do not find significant changes by only fixing the wind speed except for the SRON – Kayrros case, where discrepancy is reduced.

Common retrieval

SRON – IUP-UB shows similar wind speed values. Therefore, if we fix the retrieval for these pairs, the remaining discrepancies can only be explained by the quantification. After fixing

the retrieval, we observed only a subtle improvement. Since large discrepancy remains, we can say that most discrepancy comes from the quantification method.

For the rest of the cases, we see that large discrepancy still remains after fixing the retrieval except for the SRON – UPV case, where differences are considerably reduced. This is consistent with the previously carried out retrieval analysis, where we observed that SRON plume pixel values for the EMIT case were significantly higher than those of UPV.

Common retrievals and wind speed values
At this point, differences can only arise from the quantification process. We can observe:

- Even though SRON and Kayrros showed consistency after fixing the wind speed, we now find large disagreement after fixing both the retrieval and the wind speed.
- Large underestimation of IUP-UB in reference to the rest of the groups, although the bias is slightly reduced when fixing wind speed and retrievals. As previously discussed, this underestimation is probably due to the specific selection of parameters within the CSF method.
- Larger values estimated by Kayrros compared to the rest of the groups.
- There is an approximate agreement between SRON and UPV, where the discrepancy was mainly solved by fixing the retrieval. At this point, differences with Kayrros are probably due to differences in the U_{eff} calibration and plume masking.

Overall summary and conclusions

Overall, we can summarize the 3 cases as follows:

- For the EnMAP case, the main source of discrepancy was the retrieval.
- For the PRISMA case, plume masking had a significant impact for groups that used the IME method. Initially, UPV and Kayrros showed high agreement. Then, after using similar plume masks for UPV and SRON, estimates were consistent for these 2 groups and IUP-UB.
- For the EMIT case, an IUP-UB analysis indicated that there is a high variability in their results because of the parameter selection within the CSF method. Between UPV and SRON, we observed that retrieval was the main source of discrepancy. Remaining differences with the Kayrros group are probably coming from plume masking.

In addition, we extract the following conclusions:

- There can be differences even when using the same quantification method due to the decisions taken within the method itself. Regarding the IME method, the selection of the plume mask plays an important role in the flux estimate. Similarly, the plume coverage extent is an important factor in the CSF method.
- For the PRISMA and EnMAP cases, we found consistency between estimates derived from different quantification methods.
- There are generally non-negligible changes after fixing the retrieval. As observed in Sect. 5.4.1, there are systematic differences among groups at the retrieval level, which will cause a fraction of the flux estimate discrepancies.
- Generally, we did not see wind speed as a major contributor to discrepancy, since in these 3 cases we only observed differences up to ~ 1 m/s, while differences up to ~ 2 m/s were found in the cases from Sect. 5.3. This illustrates how the different criteria

from groups can have an impact on the final wind speed values even when extracting the values from the same product.

- Most major discrepancies have been explained and attributed, showing the success of the analysis.

6. Conclusions and further research

Plume identification and attribution to source

Regarding the emission detection performance, we found agreement among groups under advantageous observing conditions, where differences were only found in blob-like plumes at background level. The advantages of machine learning-based models in detection could not be correctly assessed due to the low number of samples in our tests. Under less ideal scenarios, emission detection seems to be highly dependent on the group criteria. A more conservative set of criteria generally leads to a lower number of detections.

Our analysis shows that groups specific criteria for plume delineation can lead to important differences, which in turn has a significant impact in quantification. Some of the different criteria considerations are whether to allow non-connected clusters to be part of the same plume, separating connected plume masks from different sources into individual masks, and human-derived errors linked to distinguish background from plume pixels. Differences in plume masks do not have to lead in differences in emission rates as long as the masking is consistent with the approach used in each group's calibration.

As for the emission attribution to source location, we generally find offsets of 2 to 3 pixels for both EMIT and EnMAP instruments. This attribution is mainly based on the groups orthorectified methane maps along to high resolution imagery, where potential emission sources can be identified. Potential shifts in the orthorectified maps and different group criteria to determine which surface elements match with the emission source (and whether only locations within the plume masks can be selected) are likely to be the main sources of discrepancy.

Assessment of methane concentration retrievals use for detection

All the groups detection retrievals aligned with the scene and instrument features, showing consistency. Moreover, we found that the selection of the retrieval spectral window has an important impact in detectability when applying matched filter methods. Combo-MF, a matched filter flavor, exploits this point reducing the occurrence of retrieval artifacts, which is helpful to filter them out of the plume mask. This is especially important for the IME quantification, since plume delineation is needed. We also found that other factors such as radiance quality filtering can lead to significant changes, since 'bad pixels' interfere with scene-derived parameters used by the matched filter methods.

General methane emission quantification assessment

The selection of the wind speed value has a great impact on the quantification results. In some cases, we found that groups selected wind speed values that differ up to 2 m/s for the same plume case, which shows potential high disagreement. These differences arise from the specific selection of values from GEOS-FP or ERA5 as data source, and the corresponding postprocessing.

Regarding the Q values reported by the teams, we also observe variability that cannot only be explained by wind speed variability. The use of different methane retrievals and quantification also play an important role. In some cases, we find remarkable differences, such as groups having a Q 8 times higher than another. In addition, there is a trend of IUP-UB showing systematically lower Q values than the rest of the groups for most cases. This point needs further research. Nevertheless, except for a few cases, we generally find that group estimates agree within the uncertainty bars.

Potential further research:

- We still cannot explain why most IUP-UB quantification results were systematically lower than those from the rest of the groups. Further research is needed to understand this point.

Detailed methane emission quantification assessment

Several aspects were investigated in more detail focusing on three individual plume observations. As for the methane retrievals used for quantification, we observed that some groups exhibit different levels of retrieval plume values and we found that most discrepancies were caused by the selection of the unit methane absorption spectrum (K spectra), which plays an important role in the matched filter performance. The methane concentration enhancement range used to fit this spectrum and the use of radiometric spectra corrections, i.e. corrections to the convolved ISRF spectra (Gorroño et al., 2023), explain a large fraction of these differences.

After a comprehensive analysis of the emission estimates of the three plume cases, we found that there can be large differences even when using the same quantification method due to the decisions taken within the method itself. Regarding the IME method, the selection of the plume mask plays an important role in the flux estimate. Similarly, the plume extent considered is an important factor in the CSF method. We also found that there can be consistency between estimates coming from different quantification methods, such as in the PRISMA and EnMAP cases after fixing the retrieval and wind speed. Moreover, there are generally non-negligible changes after fixing the retrieval, which matches with the different selection of the unit methane absorption spectra from groups. In addition, we did not see wind speed as a major contributor to discrepancy, since in these three cases we only observed differences up to ~ 1 m/s, while differences up to ~ 2 m/s were found in other cases from Sect. 5.3. Overall, most major discrepancy factors have been detected, which reveals the success of the analysis.

Potential further research:

- More in-depth analysis on how the K is extracted. Consideration of Radiative Transfer Models and how the LookUp Tables are built might explain remaining differences.
- Carry out a comprehensive analysis on cases where there is a higher difference in used wind speed values, such as the PRISMA – China case (~ 2 m/s, see Sect. 5.3), because the highest wind speed difference in the cases studied was around 1 m/s. In this manner, we would be able to assess at what extent the wind speed has an impact on the quantification results.
- Carry out a comprehensive analysis on a significantly larger number of cases to better assess what the main sources of discrepancy are.
- Link the conclusions from what was obtained in this work package to others, such as WP200, where comparisons of flux rates are also carried out and results are contrasted with those obtained using other instruments and true emission rates for controlled release experiments.

7. Changes in reference to the AIP

We followed the Algorithm Intercomparison Plan (AIP) as much as possible to carry out the intercomparisons among groups and corresponding analysis documented in this text. However, some changes needed to be made in reference to what was written in the AIP to correct, extend, and update algorithm descriptions and to enhance the analysis of the results. The main changes are:

- Groups added a detailed description of how the U_{10} is extracted. Moreover, they updated their plume delineation method and how they carry out the orthorectification for the EMIT and EnMAP data.
- Some groups updated a new version of their algorithms, which is applied in the 'Detailed methane emission quantification assessment' section.
- The use of binary masks to extract the count of methane plume pixel candidates proposed in the AIP is removed, since this cannot be correctly applied to those groups using the Combo-MF method
- According to the AIP, the retrievals shared for the 'comprehensive quantification assessment' intercomparison would be the retrievals obtained by the original UPV algorithm output. Instead, UPV shared a retrieval following Guanter et al., (2021). There are 2 advantages regarding this decision:
    - We can also evaluate a change in reference to the UPV algorithm, which can be beneficial for the analysis.
    - In case we need to understand differences in reference to the shared retrieval, it would be easier since this retrieval is less complex than the one used by UPV.
- We added some additional intercomparison figures to enhance the analysis such as the one related to the distance in estimated source location between groups and the flux rate ratio matrix.

8. References

Borchardt, J., S. J. Harris, J. M. Hacker, M. Lunt, S. Krautwurst, M. Bai, H. Bösch, H. Bovensmann, J. P. Burrows, S. Chakravarty, R. A. Field, K. Gerilowski, O. Huhs, W. Junkermann, B. F. J. Kelly, M. Kumm, W. Lieff, A. McGrath, A. Murphy, J. Schindewolf, and J. Thoböll, Insights into Elevated Methane Emissions from an Australian Open-Cut Coal Mine Using Two Independent Airborne Techniques, Environmental Science & Technology Letters, DOI: 10.1021/acs.estlett.4c01063, https://doi.org/10.1021/acs.estlett.4c01063, 2025.

Bösch, H., and M. Hilker, Inventory of point source emissions of CH4 estimated from high-resolution satellite data, Technical Report EU HORIZON-CL5-2022-D1-02 project EYE-CLIMA, deliverable 1.4, version 1.0, pp. 33, 31.12.2024, available from https://eyeclima.eu/products/public-reports/, 2024.

Carvalho, D. (2019). An assessment of NASA's GMAO MERRA-2 reanalysis surface winds. Journal of Climate, 32(23), 8261-8281.

Fuentes Andrade, B., Buchwitz, M., Reuter, M., Bovensmann, H., Richter, A., Boesch, H., and Burrows, J. P.: A method for estimating localized CO2 emissions from co-located satellite XCO2 and NO2 images, Atmos. Meas. Tech., 17, 1145–1173, https://doi.org/10.5194/amt-17-1145-2024, 2024.

Guanter, L., Irakulis-Loitxate, I., Gorroño, J., Sánchez-García, E., Cusworth, D. H., Varon, D. J., ... & Colombo, R. (2021). Mapping methane point emissions with the PRISMA spaceborne imaging spectrometer. Remote Sensing of Environment, 265, 112671.

Guanter, L., Roger, J., Sharma, S., Valverde, A., Irakulis-Loitxate, I., Gorroño, J., ... & Zavala-Araiza, D. (2024). Multisatellite data depicts a record-breaking methane leak from a well blowout. Environmental Science & Technology Letters, 11(8), 825-830.

Gordon, I. E., Rothman, L. S., Hargreaves, R. J., Hashemi, R., Karlovets, E. V., Skinner, F. M., Conway, E. K., Hill, C., Kochanov, R. V., Tan, Y., Wcisło, P., Finenko, A. A., Nelson, K., Bernath, P. F., Birk, M., Boudon, V., Campargue, A., Chance, K. V., Coustenis, A., Drouin, B. J., et al. (2022). The HITRAN2020 molecular spectroscopic database. Journal of Quantitative Spectroscopy and Radiative Transfer, 277, 107949. https://doi.org/10.1016/j.jqsrt.2021.107949

Gorroño, J., Varon, D. J., Irakulis-Loitxate, I., & Guanter, L. (2023). Understanding the potential of Sentinel-2 for monitoring methane point emissions. Atmospheric Measurement Techniques, 16(1), 89-107.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, ERA Report Series, https://doi.org/10.21957/tkic6g3wm, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloy- aux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. Roy. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Kneizys, F. X., E. P. Shettle, L. W. Abreu, J. H. Chetwynd Jr., G. P. Anderson, W. O. Gallery, J. E. A. Selby, and S. A. Clouth, 1988:User's Guide to LOWTRAN 7. Tech. Rep. 88-0177, Air Force Geophysical Laboratory, Hanscom Air Force Base, 137 pp.

Kuhlmann, G., Brunner, D., Broquet, G., and Meijer, Y.: Quantifying CO2 emissions of a city with the Copernicus Anthropogenic CO2 Monitoring satellite mission, Atmos. Meas. Tech., 13, 6733–6754, https://doi.org/10.5194/amt-13-6733-2020, 2020.

Kuhlmann, G., Koene, E., Meier, S., Santaren, D., Broquet, G., Chevallier, F., Hakkarainen, J., Nurmela, J., Amorós, L., Tamminen, J., and Brunner, D.: The ddeq Python library for point source quantification from remote sensing images (version 1.0), Geosci. Model Dev., 17, 4773–4789, https://doi.org/10.5194/gmd-17-4773-2024, 2024.

Maasakkers, J. D., Varon, D. J., Elfarsdóttir, A., McKeever, J., Jervis, D., Mahapatra, G., ... & Aben, I. (2022). Using satellites to uncover large methane emissions from landfills. Science Advances, 8(31), eabn9683.

Molod, A., Takacs, L., Suarez, M., Bacmeister, J., Song, I.-S., Eichmann, A., 2012. The GEOS-5 Atmospheric General Circulation Model: Mean Climate and Development from MERRA to Fortuna (https://portal.nccs.nasa.gov/datashare/gmao/geos-fp/das/, nASA TM-2012-104606, 2012).

Pei, Z., Han, G., Mao, H., Chen, C., Shi, T., Yang, K., ... & Gong, W. (2023). Improving quantification of methane point source emissions from imaging spectroscopy. Remote Sensing of Environment, 295, 113652.

Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., Bovensmann, H., and Burrows, J. P.: Towards monitoring localized CO2 emissions from space: co-located regional CO2 and NO2 enhancements observed by the OCO-2 and S5P satellites, Atmos. Chem. Phys., https://www.atmos-chem-phys.net/19/9371/2019/, 19, 9371-9383, 2019.

Roger, J., Guanter, L., Gorroño, J., and Irakulis-Loitxate, I.: Exploiting the entire near-infrared spectral range to improve the detection of methane plumes with high-resolution imaging spectrometers, Atmos. Meas. Tech., 17, 1333–1346, https://doi.org/10.5194/amt-17-1333-2024, 2024.

Schneising, O., Buchwitz, M., Reuter, M., Bovensmann, H., and Burrows, J. P.: Severe Californian wildfires in November 2018 observed from space: the carbon monoxide perspective, Atmos. Chem. Phys., 20, 3317-3332, https://doi.org/10.5194/acp-20-3317-2020, 2020a.

Schneising, O., Buchwitz, M., Reuter, M., Vanselow, S., Bovensmann, H., and Burrows, J. P.: Remote sensing of methane leakage from natural gas and petroleum systems revisited, Atmos. Chem. Phys., 20, 9169-9182, https://doi.org/10.5194/acp-20-9169-2020, 2020b.

Schneising, O., Buchwitz, M., Reuter, M., Weimer, M., Bovensmann, H., Burrows, J. P., and Bösch, H.: Towards a sector-specific CO/CO2 emission ratio: Satellite-based observation of CO release from steel production in Germany, Atmos. Chem. Phys., 24, 7609–7621, https://doi.org/10.5194/acp-24-7609-2024, 2024.

Sherwin, E. D., Rutherford, J. S., Chen, Y., Aminfard, S., Kort, E. A., Jackson, R. B., & Brandt, A. R. (2023). Single-blind validation of space-based point-source detection and quantification of onshore methane emissions. Scientific Reports, 13(1), 3836.

Sherwin, E. D., El Abbadi, S. H., Burdeau, P. M., Zhang, Z., Chen, Z., Rutherford, J. S., ... & Brandt, A. R. (2024). Single-blind test of nine methane-sensing satellite systems from three continents. Atmospheric Measurement Techniques, 17(2), 765-782.

Thompson, D. R., Thorpe, A. K., Frankenberg, C., Green, R. O., Duren, R. M., Aubrey, A. D., ... & Ungar, S. (2016, December). The Aliso Canyon Super-Emitter: Initial Results of observations by AVIRIS-C and the Hyperion spacecraft, with Implications for Global Spectroscopic CH 4 Monitoring. In AGU Fall Meeting Abstracts (Vol. 2016, pp. GC52A-04).

Thompson, D. R., Green, R. O., Bradley, C., Brodrick, P. G., Mahowald, N., Dor, E. B., ... & Zandbergen, S. (2024). On-orbit calibration and performance of the EMIT imaging spectrometer. Remote Sensing of Environment, 303, 113986.

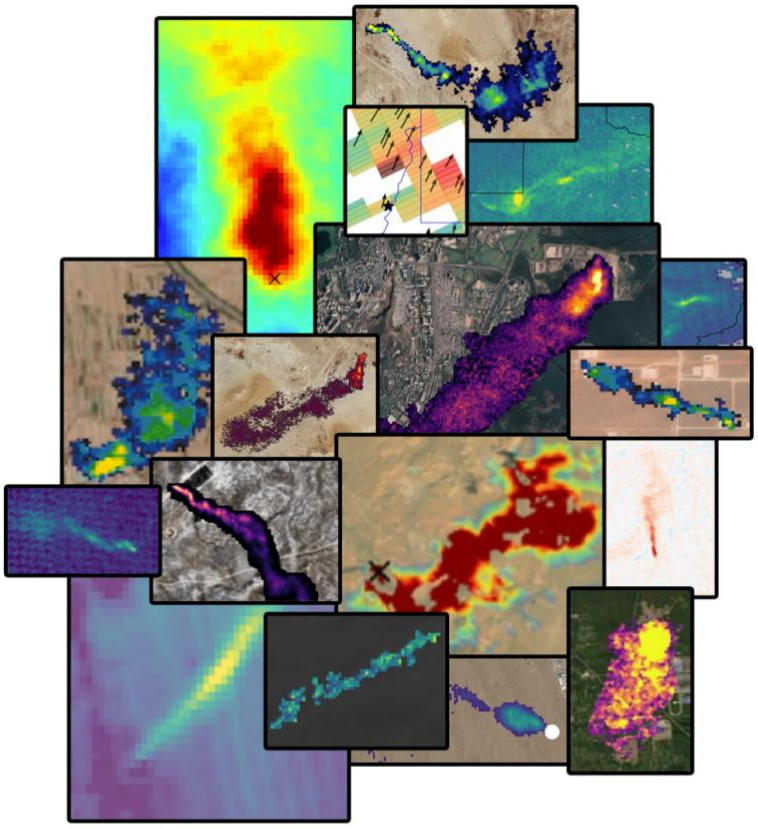
Thorpe, A. K., Green, R. O., Thompson, D. R., Brodrick, P. G., Chapman, J. W., Elder, C. D., ... & Miller, C. E. (2023). Attribution of individual methane and carbon dioxide emission sources using EMIT observations from space. Science advances, 9(46), eadh2391.

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O., Xia, Y., & Huang, Y. (2018). Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes. Atmospheric Measurement Techniques, 11(10), 5673-5686.

Zhang, X., Maasakkers, J. D., Roger, J., Guanter, L., Sharma, S., Lama, S., ... & Aben, I. (2024). Global identification of solid waste methane super emitters using hyperspectral satellites. Environmental Science & Technology.

ANNEX C (BI comparisons) to Algorithm Intercomparison Report (AIR)

WP320 – Annex for Deliverable D3.2



Authors

Lead author:

- Julian Akani-Guéry: BI (Band Imager) sections lead

All authors:

Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, Spain (UPV)
- Javier Gorrono
- Shanyu Zhou

Kayrros, Paris, France
- Julian Akani-Guéry
- Maxence Nevoret

Table of Contents

1. Executive summary

This document is an ANNEX to the Algorithm Intercomparison Report (AIR) of ESA project MEDUSA (Methane Emissions Detection Using Satellites Assessment).

The objective of MEDUSA is to develop the techniques needed for a pre-operational system able to harmonize and integrate global information on subnational (e.g., urban areas) to facility scale anthropogenic methane emissions derived using diverse satellite instruments and algorithms.

This document is ANNEX C to AIR focusing on intercomparison results for Band Imagers (BI). Within MEDUSA algorithm intercomparisons are grouped into three groups depending on satellite instrument type. These three instrument types are:

- Flux Mapper (FM): These are moderate spatial resolution (several kilometer) sensors.
- Hyperspectral Imager (HI): PRISMA, EnMAP and EMIT/ISS.
- Band Imager (BI): Sentinel-2 (S2), Landsat-8 (L8), Landsat-9 (L9) and GOES satellites.

The algorithm intercomparisons have been carried out as described in the Algorithm Intercomparison Plan (AIP, Buchwitz et al., 2024). As described in the AIP, the BI intercomparisons will focus mainly on the high-resolution constellations namely Sentinel-2 and Landsat-8/9 and will compare them in both continuous monitoring and single observation settings. The first approach intends to compare the two algorithms in a close-to-operational setting with aggregated metrics, while the second approach will include more detailed analysis on a per-scene basis. For GOES, a couple of known events will be used as case studies.

2. Overview satellite sensors

This section provides a short overview of the satellite sensors relevant for the MEDUSA BI algorithm intercomparison. These are sensors where more than one methane emission inversion algorithm is used and evaluated within MEDUSA. This is a sub-set of the satellite sensors used within MEDUSA.

2.1. Sentinel-2 (S2 series)

Sentinel-2 is a multispectral imaging mission operated by ESA with three satellites: Sentinel-2A launched in June 2015, Sentinel-2B launched in March 2017, and Sentinel-C in June 2024. While Sentinel-2A was originally scheduled for decommissioning and replacement by Sentinel-2C, an exceptional one-year extension campaign began in March 2025. As our analysis focuses on imagery before June 2024, Sentinel-2C is excluded from this study. Each of them carries a push-broom instrument measuring 13 spectral bands: 4 VNIR bands at 10 m spatial resolution, 6 SWIR bands at 20 m, and 3 bands at 60m for atmospheric correction purposes and cloud screening. The combination of the two satellites (Sentinel-2A and Sentinel-2B in our case) provides global coverage with a 5-day revisit time.

2.2. Landsat-8 (L8) and Landsat-9 (L9)

Landsat 8 and 9 are part of NASA's Landsat program and were respectively launched in February 2013 and September 2021. They carry a push-broom imager measuring 9 spectral bands between 433 and 2300 nm: 5 VNIR and 3 SWIR bands at 30 m spatial resolution, and 1 panchromatic band at 15 m spatial resolution. The combination of the two satellites provides global coverage in an 8-day revisit time.

2.3. GOES

GOES 3rd generation series gathers four geostationary satellites developed by NOAA and NASA.

- GOES-R (or GOES-16) was launched in November 2016 and is located above Central America and the Eastern US.
- GOES-S (or GOES-17) was launched in March 2018 and is located above the Pacific Ocean.
- GOES-T (or GOES-18) was launched in March 2022 and is located above Central America and the Western US.
- GOES-U (or GOES-19) was launched in June 2024 and is located above the Atlantic Ocean.

These satellites carry among other instruments an Advanced Baseline Imager (ABI), which is a multispectral sensor. It measures 2 visible bands at 500 m spatial resolution at nadir, 3 short wave infrared (SWIR) bands at 1 km and 11 medium wave infrared (MWIR) bands at 2 km.

Only GOES-R imagery is used in this study.

GOES-R routinely operates following two acquisition modes:

- The CONUS (Continental United States) mode corresponds to the acquisition of an image every 5 minutes on the CONUS sector defined by the US, and parts of Canada and Mexico.
- The Full Disk mode corresponds to the acquisition of an image every 15 minutes on the Earth disk visible from the satellite.

These satellites offer unique temporal resolution compared to Low Orbiting satellites.

3. Algorithm description

This section provides a brief overview of methane plume detection and emission rate estimation algorithms as will be used and assessed within the MEDUSA BI algorithm intercomparison task.

In this document, algorithms to obtain methane emission information are also referred to as Level 4 (L4) algorithms. Input are lower-level data products such as Level 2 (atmospheric methane information from and for individual satellite footprints or ground pixel) or Level 1 (i.e., radiance) products. For all band imagers, the starting point is Level 1 products.

3.1. Kayrros algorithm for Sentinel-2, Landsat-8/9, GOES

These Band Imagers (BIs) have two band measurements in the SWIR domain: a band Bstrong where methane strongly absorbs (around 2200 nm) and a band Bweak where methane absorbs much less (around 1600 nm). Kayrros algorithm leverages correlation between these two bands for other absorbing materials to enhance methane by computing the band ratio Bweak / Bstrong. To further enhance methane, images from previous observations are used to estimate and remove the background. Only previous observations are used since the algorithm is running on a live basis as soon as the satellite image is available. The images obtained after background removal are called methane residuals.

For Sentinel-2 and Landsat-8/9 (Ehret et al., 2022), the background for a given date is estimated as its linear regression over previous dates. The linear regression is done on log (Bweak / Bstrong) rather than Bweak / Bstrong to reduce the influence of strong albedo variations that may end up leading the regression. The 10 most recent observations of the same location with less than 10 % of cloud cover are used as reference images in the regression.

For GOES (Groshenry et al., 2024), the background on a date D at a timestamp T is estimated as the median of band ratios at timestamps T-5', T, T+5' on previous dates D-N, D-(N-1), …, D-1. In practice, N is set to 10. Using surrounding timestamps allows to have similar environmental conditions as the target observation. This is particularly important for the solar zenith angle, which is influencing the illumination, the shadows, and the retrieval noise. The Clear Sky Mask product (Heidinger et al., 2012) is used to mask clouds in background estimation.

The detection and plume delineation are performed in an automated way by a deep learning model on the methane residuals. Such an algorithm is still under development for GOES so it is done manually for the time being.

Every Sentinel-2 and Landsat 8/9 detection is quantified using the Integrated Mass Enhancement (IME) method with an effective wind speed Ueff = 0.33 * U10 + 0.45 (Varon et al., 2021). The ERA5 10m wind speed reanalysis data (Hersbach et al., 2018) is used, assuming a default 0.5 m/s uncertainty on each component of the wind vector. A flux rate uncertainty is estimated by combining the uncertainty on the wind data and an empirical 1-σ estimate of the retrieval noise.

For GOES, the IME is computed for each timestamp using the binary mask computed for each frame. The mass of emitted methane is directly inferred by computing the median from the time series of IME results.

3.2. UPV algorithm for Sentinel-2A&B, Landsat-8/9, GOES

The retrieval of methane concentration enhancement for hyperspectral sensors typically incorporate the full SWIR spectral region and fit those observations against a modelled radiance (Thorpe et al., 2014; Jacob et al., 2016). In the case of multispectral instruments the SWIR spectral region is not fully resolved and, although a data-driven method is possible, the selected method involves a ratio of bands together with temporal normalisation.

This methane enhancement transmittance is approximated as the ratio between the radiance at a spectral channel affected by methane absorption (this is B12 for Sentinel 2 and B7 for Landsat) and a methane-free reference band (this is B11 for Sentinel 2 and B6 for Landsat). Thus, the proposed rationale to detect methane plumes for the S2/Landsat mission is based on an estimation of the plume transmission defined by the ratio:

$$\frac{\rho}{\rho_{\text{ref}}} = \frac{\rho_{\text{B12plume}} / \rho_{\text{B11plume}}}{\rho_{\text{B12plumefree}} / \rho_{\text{B11plumefree}}}$$

where the subscript plume and plumefree refer to the overpasses with methane plume emissions and without it, respectively, and ρ refers to the TOA reflectance. Note that here we refer to B12 and B11 from Sentinel 2 with equivalent descriptions for Landsat B7 and B6.

The selection of the temporal normalisation (i.e., plumefree) is not predefined but it is based on an ad-hoc solution. As a rule of thumb, the closest observation (or observations) with the same satellite unit and observation angle is selected. For Sentinel 2, this occurs every 10 days. This default selection is optimum for heterogeneous scenes where angular and temporal effects can have a major impact.

However, this criterion can be relaxed for a more homogeneous and isotropic surface where a mix of satellites and viewing conditions can improve the temporal normalisation.

The next step seeks to relate the transmittance image to a methane concentration enhancement ΔXCH4. Mathematically this relationship can be written as:

$$T_{\text{plume}}(\lambda) \sim \frac{L}{L_{\text{ref}}} = e^{-\text{AMF}\cdot\sigma_{\text{CH4}}\cdot\Delta X_{\text{CH4}}}$$

where L and Lref represent the radiance of the methane-sensitive band and the methane-free reference band, respectively and AMF refers to the air mass factor.

In the algorithm, this mathematical relationship between the term ΔXCH4 and Tplume(λ) is established with a radiative transfer in a look-up-table (LUT). A LUT with the relationship between these two quantities and the methane enhancement. This LUT incorporates the small but non-negligible effect of methane absorption in Sentinel 2 B11 and Landsat B6.

The methane enhancement maps ΔXCH4 are further converted into methane flux rates Q by the means of the integrated mass enhancement (IME) model (Frankenberg et al., 2016; Varon et al., 2018). The ΔXCH4 values are converted from ppm units to kg (Thompson et al., 2016; Duren et al., 2019). This is the value of IME that is computed as the sum of the ΔXCH4 pixels defined by the plume mask considering the Sentinel 2 and Landsat pixel resolution, the Avogadro's law where 1 mole of gas occupies 22.4 L, and the molar mass of methane. The emission flux rate is calculated by relating the mass enhancement to the flux rate using this expression:

$$Q = \frac{U_{\text{eff}} \cdot \text{IME} \cdot 3600}{L}$$

where L is a plume length scale in m (square root of the entire area covered by the plume pixels), and Ueff is the effective wind speed derived from WRF-LES simulations by Varon et al., 2021, and has the expression Ueff = 0.33 * U10 + 0.45.

The value of U10 is obtained from the closest pixel and a linear temporal interpolation of ERA 5 Land or GEOS-FP datasets.

4. Continuous monitoring intercomparison

This first phase of the band imagers intercomparison plan tries to replicate operational conditions, where target regions are monitored in near real time and will be limited to Sentinel-2 and Landsat. Indeed, at the difference of hyperspectral imagers, band imagers take images of the Earth continuously with a good revisit frequency, allowing to monitor methane emissions over wide areas and long periods of time.

However, band imagers suffer from a coarse spectral resolution which makes observing methane emissions in complex scenes and identifying small emissions hard. For this reason, this first phase focuses on two target regions with bright and rather simple backgrounds to maximize the probability of detections.

4.1. Target regions

The two target regions are:

- A part of the **Bugdayly gas field in Turkmenistan** (Figure 4.1, left), over the full 2023 year. Turkmenistan presents ideal conditions for using band imagers with bright desertic scenes, almost no clouds, and a lot of gas infrastructures emitting methane in the atmosphere almost continuously, making this country one of the largest methane emitters in the world. **Two target sub-regions** of roughly 10 km² containing potential emitters have been defined and are covered by a few hundred observations during the year 2023.

- A **compressor station in the Permian basin** (Figure 4.1, right) between the beginning of July 2020 and the end of September 2020. During these 3 months an incident at the compressor station caused methane leaks. This is a complementary case study to the one in Turkmenistan, since the high revisit time of band imagers is used to monitor a known source, and the content of the scene is more complex and changing. This can notably affect the plume segmentation and quantification steps.

Figure 4.1: Bugdayly gas field in Turkmenistan (left), Permian compressor station (right) (location: -102.0423, 31.7317) with heterogeneous background.

4.2. Intercomparison results for target regions

In this section, results are consistently presented by target region, as the detection capability of band imagers is highly dependent on background conditions. The Bugdayly gas field in Turkmenistan is split into two sub-regions, as shown in Figure 4.1 left, with the bottom-left sub-region designated as “TKM1”, and the upper-right sub-region “TKM2”. The compressor station in the Permian Basin is designated as “USA”.

4.2.1. Detection performance

The first comparison element between UPV and Kayrros algorithms is the detection performance. A key element here is to understand the pros and cons of an automatic detection, as done by Kayrros, against a manual approach, as done by UPV.

Sentinel-2

As shown in Figure 4.2, UPV detected 55 Sentinel-2 plumes: 10 for TKM1, 38 for TKM2 and 7 for USA whereas Kayrros detected 43 Sentinel-2 plumes: 6 for TKM1, 28 for TKM2 and 9 for USA. This case-study is in a blind detection mode, meaning emitting sources are unknown. Consequently, we are lacking ground-truth data, which prevents us from identifying false positives. In a first approach, and after a manual review, we concluded that all plumes are true positives and the recall metric was calculated accordingly by region for the two providers.

As we can see in Table 4.1, there is a significant drop in recall for Kayrros in the two regions of Turkmenistan. For TKM2, given the high number of detected plumes, it seems that the infrastructure is a constant emitting source. As explained in Table 2, Kayrros uses an automatic selection of images to build the background. By taking the closest images in the past, without manual check, it is possible to use images with a plume to build the background.

Surprisingly, the automatic approach helps identifying more plumes in the USA region, which is unexpected as the background is more complex and changing.

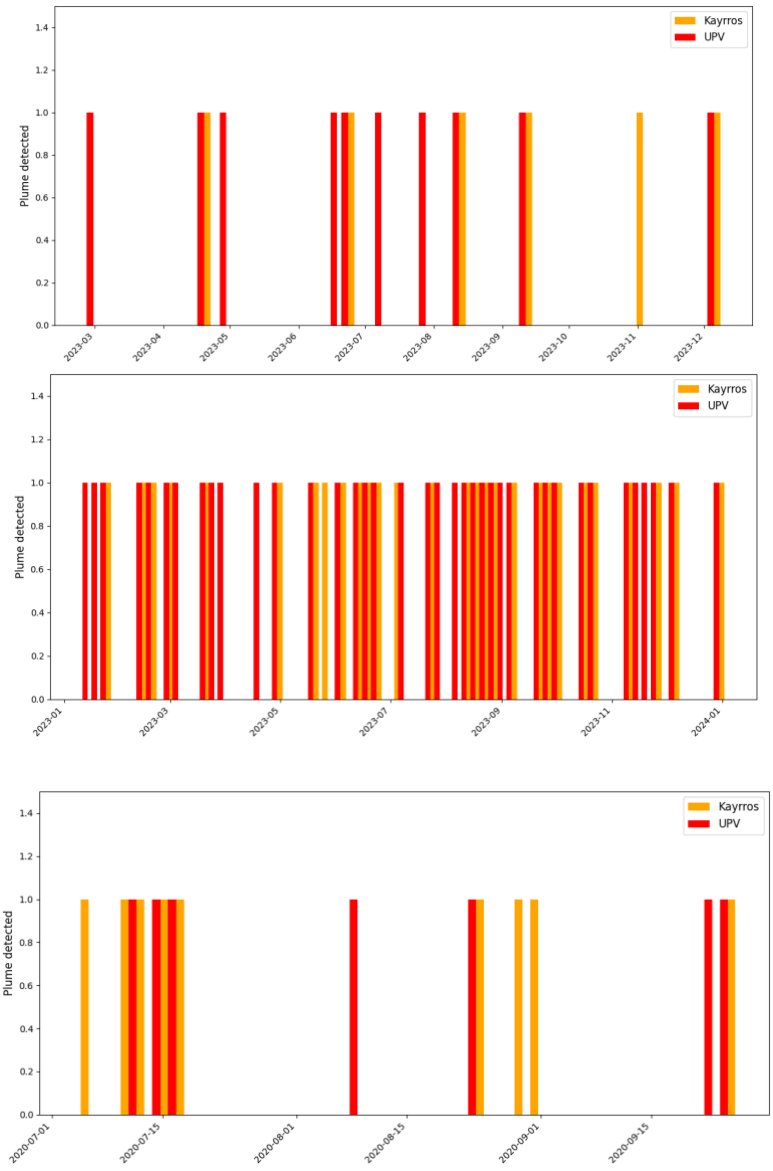


Figure 4.2: Sentinel-2 detection comparisons between Kayrros (orange) and UPV (red), for TKM1 (upper), TKM2 (middle) and USA (lower).

Region or sub-region	**Kayrros recall** "Automatic approach + manual labelling"	**UPV Recall** "Manual approach"
TKM1	0.55	0.91
TKM2	0.7	0.95
USA	0.82	0.64

Table 4.1: Sentinel-2 detection performance by region for Kayrros and UPV. The recall is equal to the number of plumes detected divided by the total number of plumes.

Another key element of comparison for the detection performance is the time spent to provide the datasets. Each data provider was asked the time spent for the two main steps of their algorithm:

- The time spent to select the images to build the background image used to obtain the methane residuals.
- The time spent to detect the plumes from the residuals, to delineate them and to pinpoint the source location.

As demonstrated in Table 2, Kayrros required 10 times less effort than UPV for dataset construction (1.5 hours versus 14.5 hours). While the manual approach delivers excellent performance with 0.89 overall recall, it is significantly more time-consuming as expected, compared to the automated approach which achieved 0.69 recall.

In operational deployment, both approaches can complement each other:

first, a deep learning detection model can rapidly identify super-emitter locations within a region of interest

- then, a manual labeling of all acquisitions over these locations.

This approach is currently followed in the IMEO-MARS platform (Vaughan et al. 2024).

	Kayrros	**UPV**
Methodology for background image selection	Automatic 10 closest images of the same satellite in the past with less than 10% coverage	Manual choice 2 closest images of the same satellite (past and future) with no cloud
Time spent for background image(s) selection (in hours)	0	4
Methodology for plume detection	Automatic by deep learning	Manual
Methodology for plume delineation	Automatic by deep learning	Manual
Methodology for source location	Manual	Manual
Time spent for plume detection, delineation and source location (in hours)	1.5	5
Total time spent	1.5	9
Recall for all three regions	0.69	0.89

Table 4.2: Sentinel-2 overall comparison of the two detection systems.

Landsat-8/9

As shown in Figure 4.3, UPV detected 27 Landsat-8/9 plumes: 4 for TKM1, 21 for TKM2 and 2 for USA whereas Kayrros detected 12 Landsat-8/9 plumes: 1 for TKM1, 7 for TKM2 and 4 for USA. Similarly to Sentinel-2, as this case study is in blind detection mode, plumes from both providers are all considered as true positives and recall metrics will be calculated accordingly.

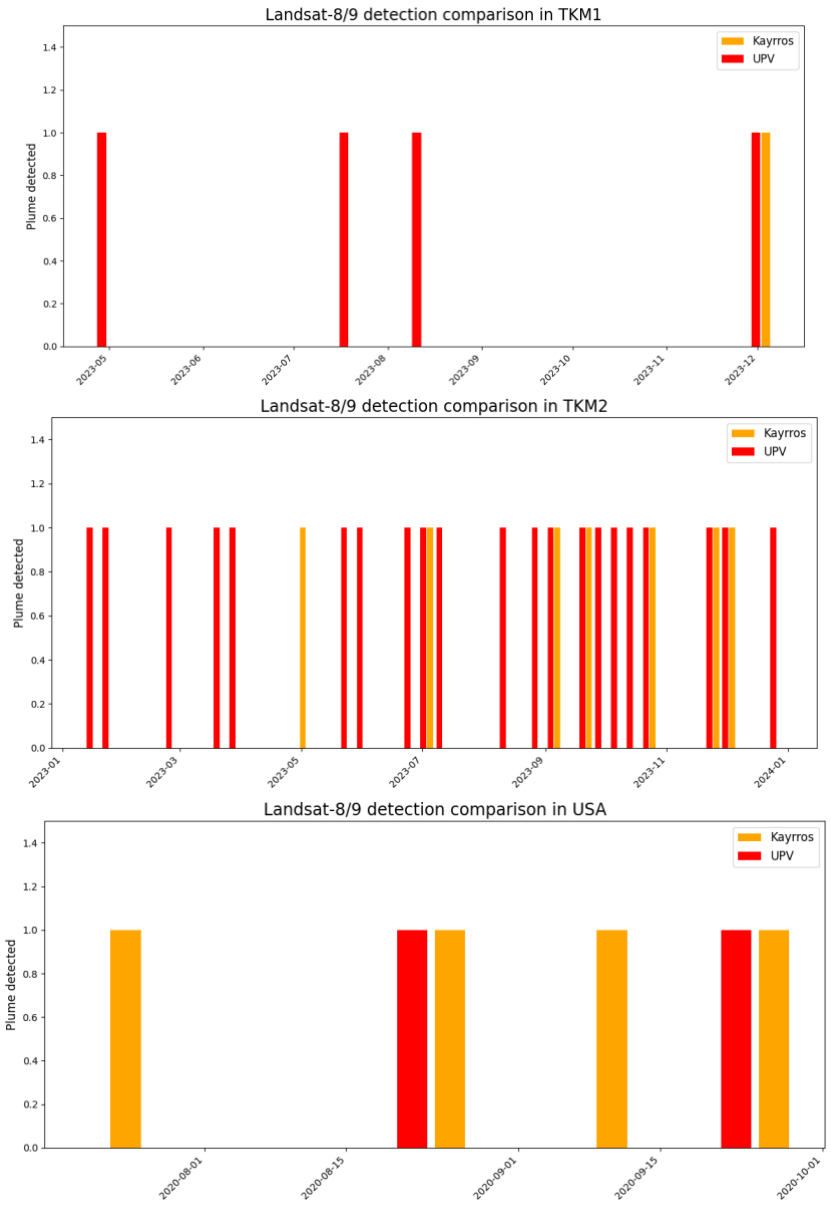


Figure 4.3: Landsat-8/9 detection comparisons between Kayrros (orange) and UPV (red), for TKM1 (upper), TKM2 (middle) and USA (lower).

Consistent with Sentinel-2 findings, Landsat-8/9 data analysis demonstrated a substantial reduction in recall performance for the Kayrros detection system across both Turkmenistan regions. To elucidate the underlying causes of false negative classifications, we performed a systematic analysis to differentiate between plumes not detected by the deep learning model and the detections human annotated as false positive.

Our analysis revealed that the deep learning detection model correctly identified 2 methane emission events in TKM1 and 4 events in TKM2; however, these detections were classified as false positives during human validation. The false positive classification resulted from a relatively low signal in the Kayrros methane retrievals, which exhibited markedly lower intensities compared to corresponding UPV measurements. This finding emphasizes the critical role of preprocessing algorithms in the overall performance of the detection pipeline.

Region or sub-region	**Kayrros recall** "Automatic approach + manual labelling"	**Kayrros detection model only** "Automatic approach only"	**UPV Recall** "Manual approach"
TKM1	0.25	0.75	1.0
TKM2	0.32	0.5	0.95
USA	1.0	1.0	0.5

Table 4.3: Landsat-8/9 detection performance by region for Kayrros and UPV. The recall is equal to the number of plumes detected divided by the total number of plumes, including the plumes that were first considered false positives.

	Kayrros	**UPV**
Methodology for background image selection	Automatic 10 closest images of the same satellite in the past with less than 10% coverage	Manual choice 2 closest images of the same satellite (past and future) with no cloud
Time spent for background image(s) selection (in hours)	0	2.5
Methodology for plume detection	Automatic by deep learning	Manual
Methodology for plume delineation	Automatic by deep learning	Manual
Methodology for source location	Manual	Manual
Time spent for plume detection, delineation and source location (in hours)	1	3
Total time spent	1	5.5
Recall for all three regions	0.4	0.9

Table 4.4: Landsat-8/9 overall comparison of the two detection systems.

4.2.2. Methane emission quantification

Sentinel-2

A significant number of plumes (36) have been identified by both UPV and Kayrros. Figure 4.4 presents point charts for each region comparing quantifications from both data providers. The emission estimates show systematic differences (UPV higher than Kayrros) but still overlap within their uncertainties.

Orthogonal Distance Regressions were performed region by region to evaluate the statistical validity of this observation. Two “outliers” were excluded at this stage due to their excessive influence on the regression: 20231205 TKM1 and 20231230 TKM2. For TKM1 and USA, regression coefficients indicate UPV quantifications are approximately 64% - 73% higher than Kayrros on average. While R^2 values are relatively good for these regions, this must be considered alongside the limited sample sizes. Notably, the coefficients for TKM1, TKM2 and USA are relatively similar (1.64 vs 1.78 vs 1.73).

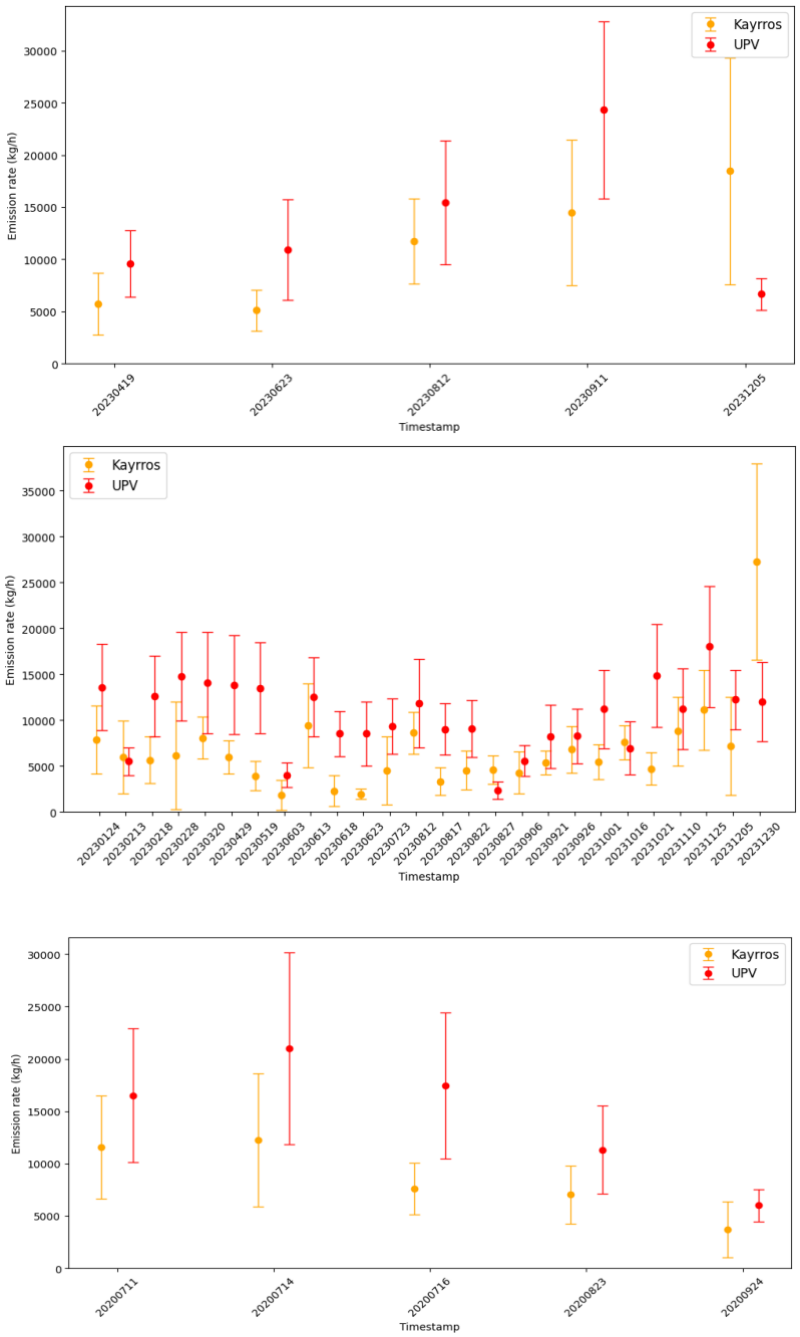


Figure 4.4: Emission rate comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes for TKM1 (upper), TKM2 (middle) and USA (lower).

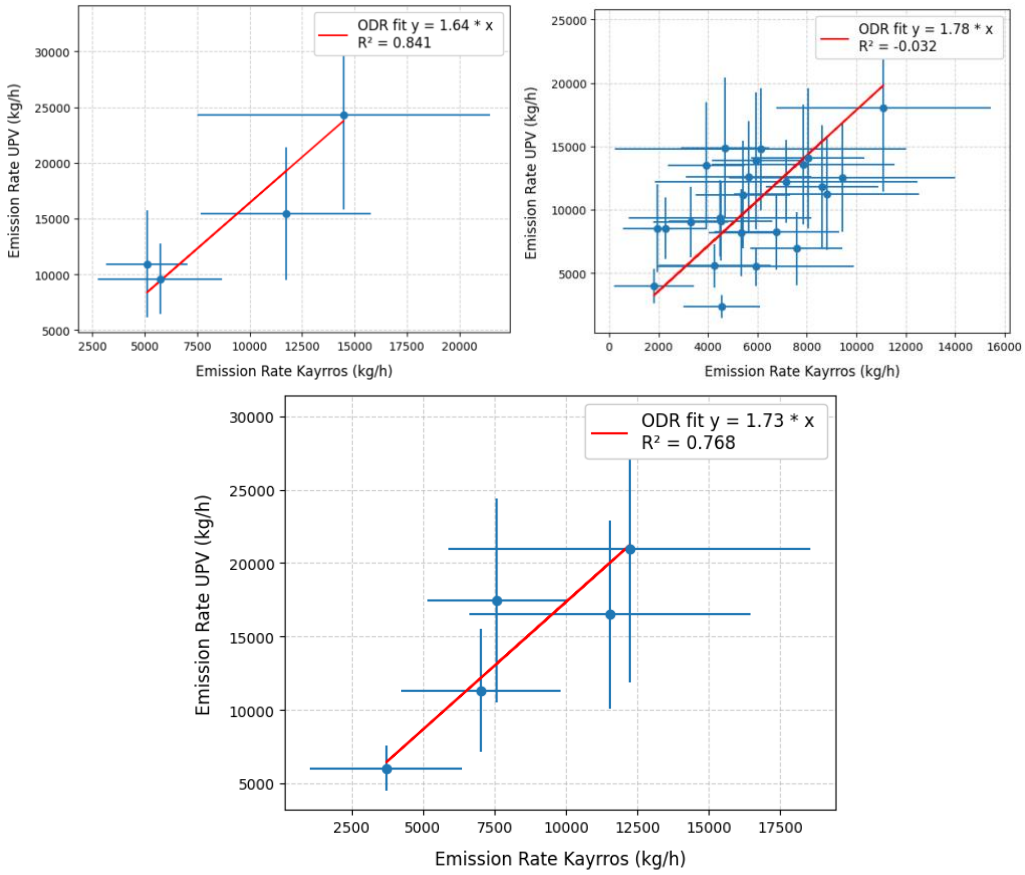


Figure 4.5: Orthogonal Distance Regression (ODR) between Kayrros and UPV Sentinel-2 emission rates for TKM1 (upper left), TKM2 (upper right) and USA (lower).

Since both Kayrros and UPV employ the IME methodology for plume emission rate quantification, the systematic bias in the IME formula can originate from either Ueff or IME/L parameters.

On the one hand, Figure 4.6 demonstrates that Ueff differences are usually minimal between the two methodologies. This outcome is anticipated since both UPV and Kayrros utilize identical wind data sources (ERA-5 Reanalysis) and apply the same effective wind calculation formula. The only methodological distinction lies in temporal handling: Kayrros uses the nearest past date for wind data, while UPV employs time interpolation. However, this interpolation approach can generate significant discrepancies, reaching up to 38% (as observed for TKM1 on June 23, 2023).

On the other hand, Figure 4.7 shows that IME/L is indeed responsible for the bias between Kayrros and UPV emission rate estimates. As the two parameters IME and L are positively correlated, it is hard to distinguish between one of the two. Figure 4.8 shows that there can be a significant difference in the plume delineation: Kayrros

mask is bigger, more detailed than the one by UPV. To properly assess the impact of the plume mask, it is necessary to use the same L2 product for the quantification. This topic will be explored in a more detailed analysis for several single observations in section 5.

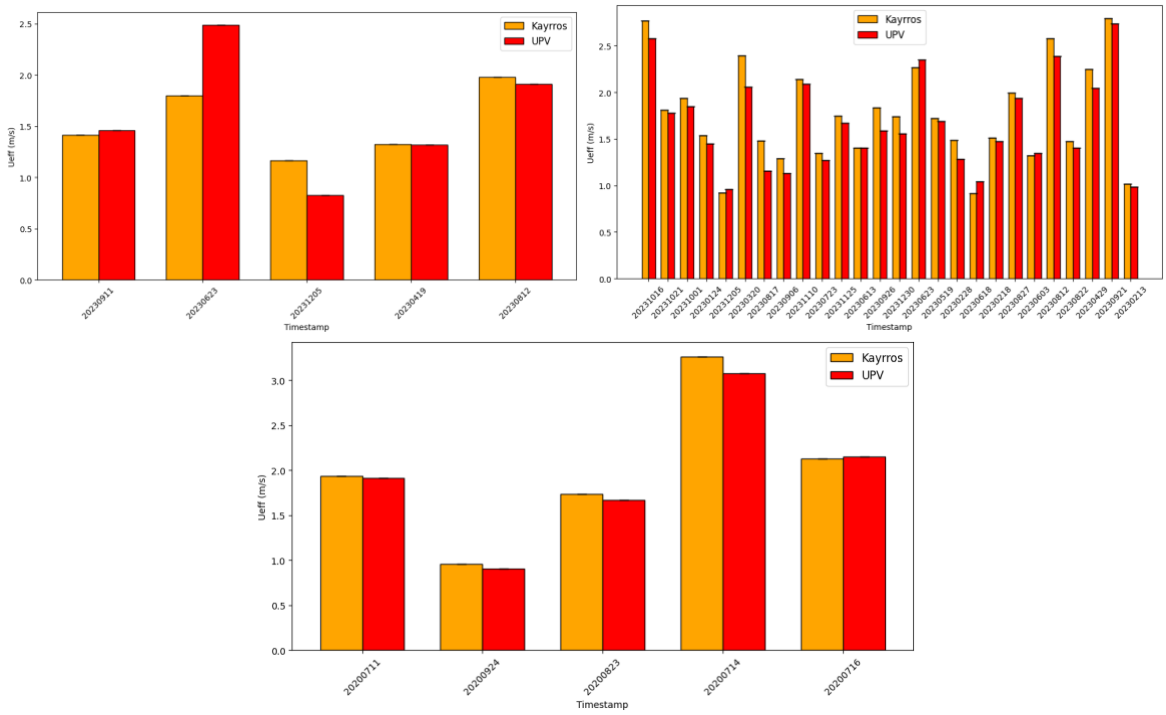


Figure 4.6: Effective wind (Ueff) comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes, for TKM1 (upper left), TKM2 (upper right) and USA (lower).

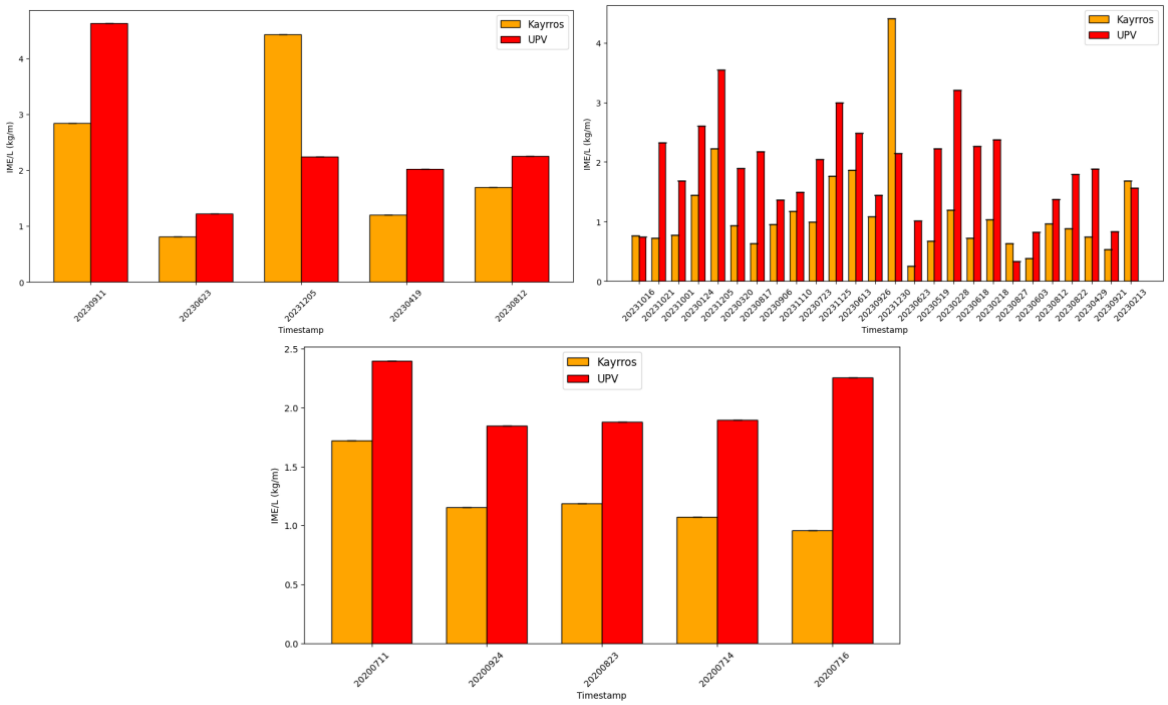


Figure 4.7 IME/L comparisons between Kayrros (orange) and UPV (red) Sentinel-2 plumes, for TKM1 (upper left), TKM2 (upper right) and USA (lower).

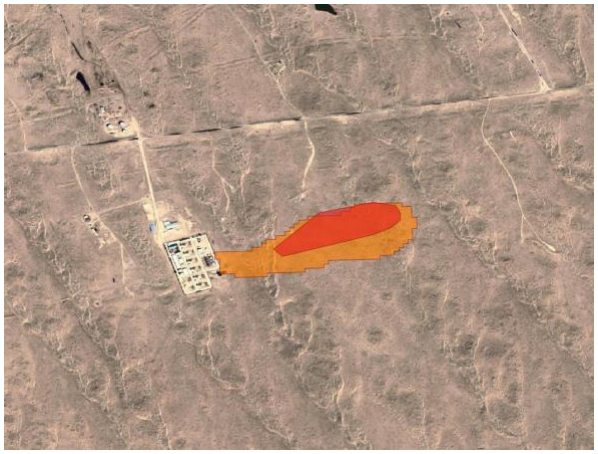


Figure 4.8: Kayrros (orange) and UPV (red) masks delineated for 20230623 Sentinel-2 plume in TKM1.

Landsat-8/9

Both UPV and Kayrros identified a limited set of 9 common plumes. Figure 4.9 displays regional comparisons of quantifications from both providers, with uncertainty intervals showing non-zero intersections in most instances. Consistent with Sentinel-2 observations, the difference (UPV - Kayrros) has a systematic positive bias.

A global Orthogonal Distance Regression analysis confirms the statistical significance of this trend. Figure 4.10 reveals strong correlation between the datasets ($R^2 = 0.909$), with the ODR coefficient indicating approximately 81% relative difference between UPV and Kayrros quantifications.

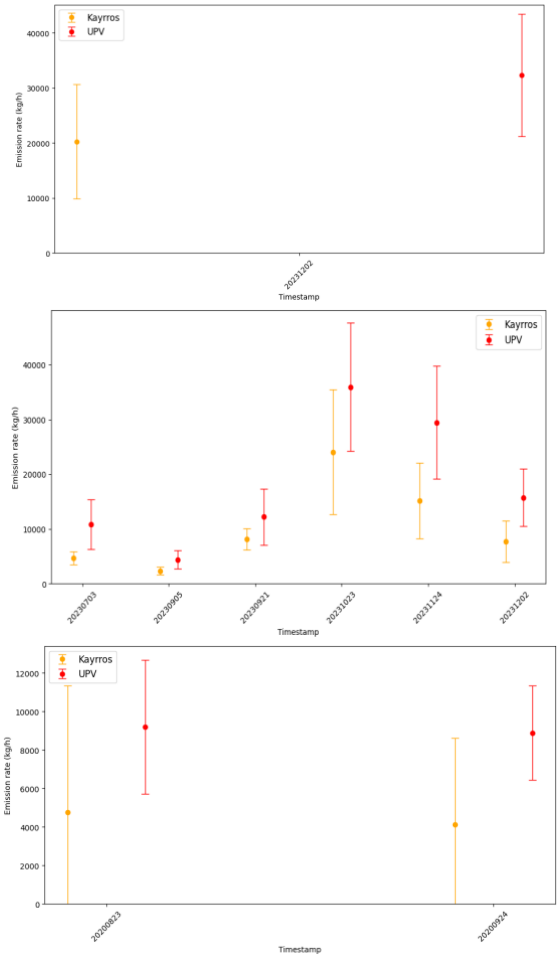


Figure 4.9: Emission rate comparisons between Kayrros (orange) and UPV (red) Landsat-8/9 plumes for TKM1 (upper), TKM2 (middle) and USA (lower).

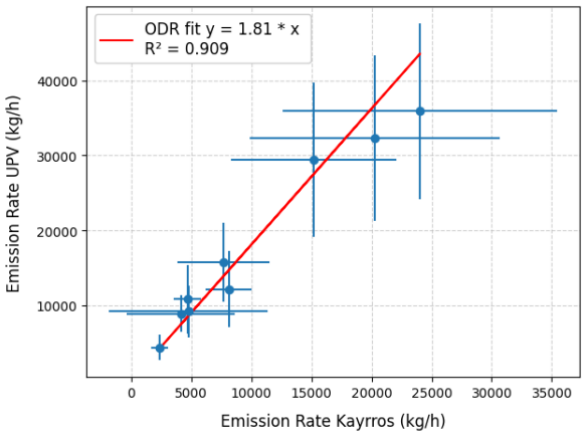


Figure 4.10: Orthogonal Distance Regression (ODR) between Kayrros and UPV Landsat-8/9 emission rates for all three regions.

Figure 4.11 reveals that relative Ueff differences between UPV and Kayrros remain minimal: under 5% in most cases, with only two dates showing larger discrepancies (maximum 14% for TKM2 on September 5, 2023).

In contrast, relative IME/L differences between Kayrros and UPV consistently hover around 50%. As with Sentinel-2 analysis, detailed examination of these IME/L discrepancies requires access to both retrievals and masks, which is conducted for selected images in Section 5.

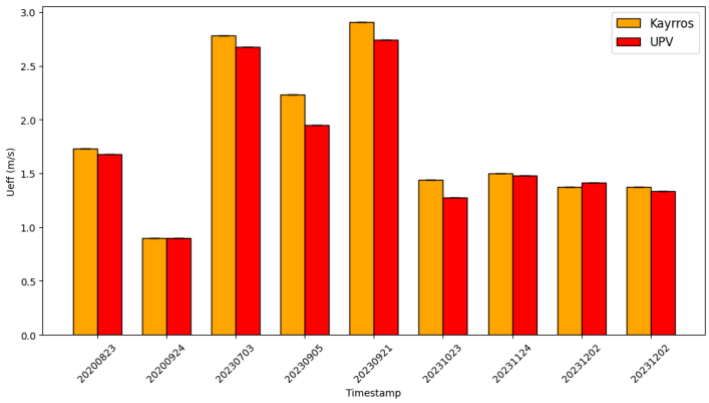


Figure 4.11: Effective wind (Ueff) comparisons between Kayrros (orange) and UPV (red) Landsat-8/9 plumes, for all three regions.

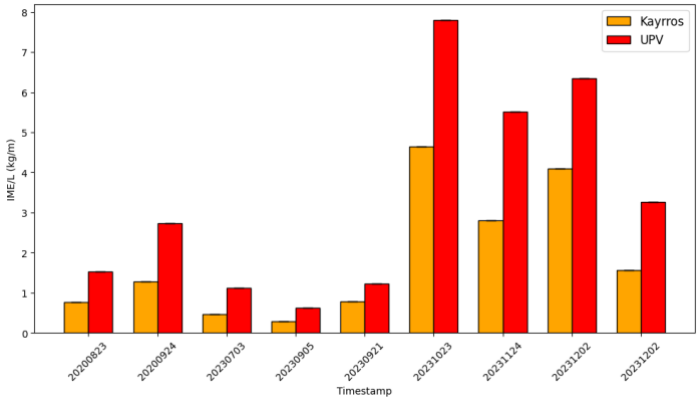


Figure 4.12: IME/L comparisons between Kayrros (orange) and UPV (red) Landsat-8/9 plumes, for all three regions.

4.2.3. Analysis of missed plume detections

Sentinel-2

This section aims to identify the root causes behind plumes detected by one provider but missed by the other. Understanding these factors is crucial for potential improvements. In deep learning applications for specific tasks like "methane plume detection," performance degradation typically stems from training set biases.

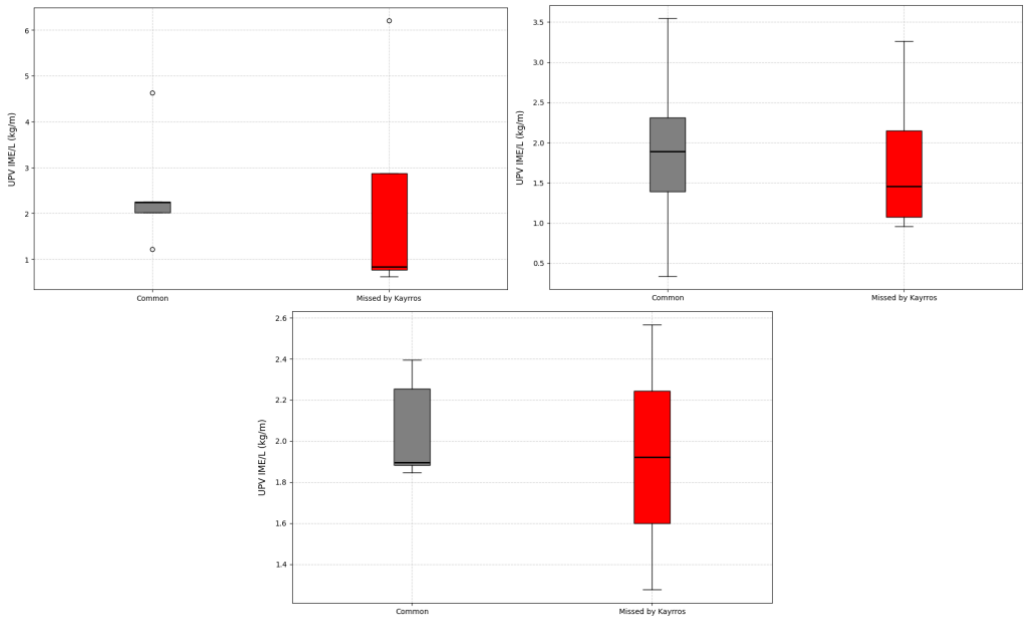


Figure 4.13: Box plots of IME/L of UPV Sentinel-2 plumes, for TKM1 (upper left), TKM2 (upper right) and USA (lower). Grey boxes represent plumes detected by both Kayrros and UPV, while red boxes show plumes detected exclusively by UPV but missed by Kayrros.

According to Figure 4.13, the IME/L distribution for UPV-detected Sentinel-2 plumes that Kayrros missed shows no clear differentiation from common plumes. Conversely, based on Figure 4.14, for Sentinel-2 plumes that UPV failed to detect in the USA, the IME/L values are significantly lower. The interpretation of these boxplots are to be put into perspective, given the limited sample size (only 2 plumes missed by UPV in the USA).

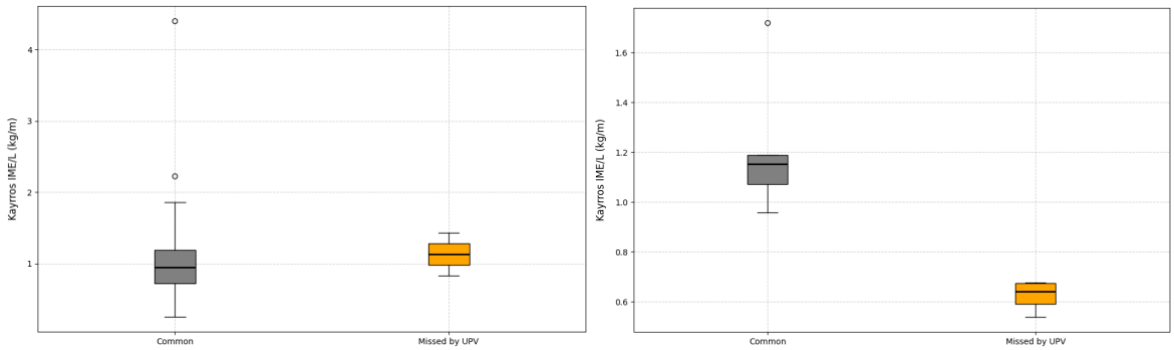


Figure 4.14: Box plots of IME/L of Kayrros Sentinel-2 plumes, for TKM2 (left) and USA (right). Grey boxes represent plumes detected by both Kayrros and UPV, while orange boxes show plumes detected exclusively by Kayrros but missed by UPV.

Landsat-8/9

Based on Figure 4.15, Landsat-8/9 plumes missed by Kayrros exhibit IME/L values in the lower ranges. However, for Ueff, as illustrated by Figure 4.16, the distributions show high similarity between common plumes and those missed by Kayrros.

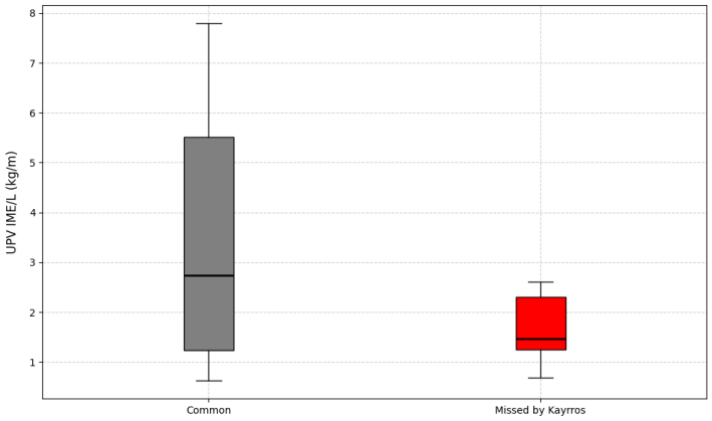


Figure 4.15: Box plots of IME/L of UPV Landsat-8/9 plumes, for all three regions. Grey boxes represent plumes detected by both Kayrros and UPV, while red boxes show plumes detected exclusively by UPV but missed by Kayrros.

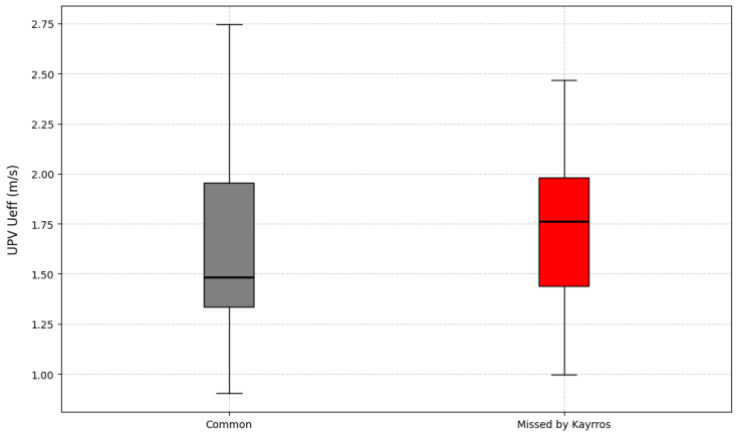


Figure 4.16: Box plots of Ueff of UPV Landsat-8/9 plumes, for all three regions. Grey boxes represent plumes detected by both Kayrros and UPV, while red boxes show plumes detected exclusively by UPV but missed by Kayrros.

4.3. Summary of comparisons for target regions

Detection performance.
For Sentinel-2, Kayrros' automated detection approach saves 85% of time, but at the cost of a 20% performance decrease compared to the manual method. The observation is relatively similar for Landsat-8/9: 82% more-time efficiency but with a more substantial performance trade-off (50% recall reduction). In operational monitoring, both approaches can be complementary: the deep learning model can be used to initially identify super-emitter location, which are then manually monitored to ensure the highest possible recall rate.

Methane emission quantification.
We observed systematic differences of 76% between UPV and Kayrros Sentinel-2 plume quantifications, and 81% for Landsat-8/9. Both providers employ the IME methodology for flow rate quantification. The discrepancy originates from either the effective wind (Ueff) or the integrated mass enhancement divided by characteristic length (IME/L):

- **Ueff**: Both providers use the same wind source and effective wind formula. The only difference is that Kayrros selects the nearest wind point from the past, while UPV applies temporal interpolation. As anticipated, this produces minimal differences (~5%), except for two cases showing 14% relative error.
- **IME/L**: This parameter accounts for most of the systematic bias in both Sentinel-2 and Landsat-8/9 quantifications. To investigate further and determine whether the mask or L2 product drives this difference, we selected individual scenes for detailed analysis in Section 5. Ground-truth data from sources like Stanford controlled release campaigns is needed for validation

Analysis of missed plume detections.
For Sentinel-2, no clear pattern emerges to explain plumes missed by either provider. For Landsat-8/9, Kayrros' neural network appears to regularly miss lower-intensity plumes (characterized by lower IME/L values).

5. Single observation intercomparison

5.1. Single observations

For high-resolution band imagers, the second phase of the BI intercomparison plan focuses on analyzing differences between the outputs of the algorithms for carefully selected observations, including identified outliers of the previous phase as well as a few additional observations from two sites (Figure 5.1) with more challenging conditions:

- **An oil and gas field in Iraq** (Figure 5.1, left). Although this scene features a mostly desertic scene, the background is much more heterogeneous and complex, notably due to the presence of many flaring stacks.

- **An oil and gas field in Venezuela** (Figure 5.1, right). Venezuela is one of the largest methane emitters in the world, but due to very challenging observation conditions, very few satellite detections of methane emissions have been made in this country. The chosen site is characterized by dense vegetation with a changing aspect throughout the year, frequent cloud coverage, and a lot of flaring. All these elements are likely to hinder the background removal step, which is essential in band imagers' methane retrieval algorithms.

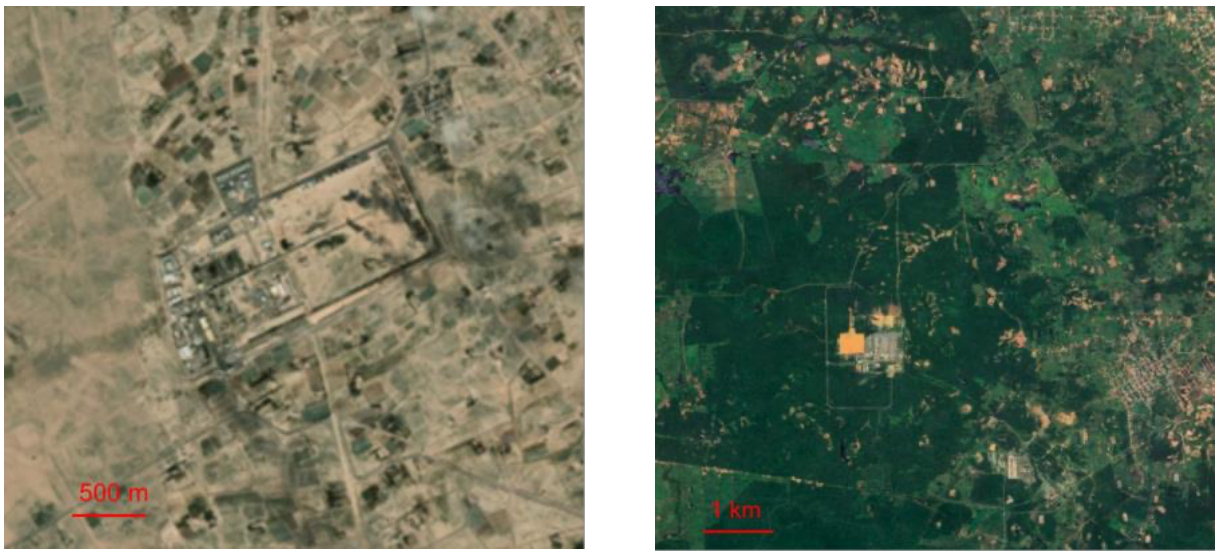


Figure 5.1: O&G field in Iraq (left) (location: 47.742532, 30.274139) and O&G field in Venezuela (right) (location: -64.532524, 9.390828).

Satellite	*Study area*	*Date*	*Comment*
Sentinel-2	*Venezuela*	*01/01/2023*	*Green vegetation, small clouds*
Sentinel-2	*Venezuela*	*22/11/2023*	*Green vegetation, small clouds*
Landsat	*Venezuela*	*29/03/2023*	*Dry vegetation and flaring*
Landsat	*Venezuela*	*05/09/2023*	*Green vegetation, small clouds*
Sentinel-2	*Iraq*	*25/01/2022*	*Bright, light flaring smokes*
Sentinel-2	*Iraq*	*04/07/2022*	*Large flaring smokes*
Landsat	*Iraq*	*03/02/2022*	*Large flaring smoke, surrounding dates are cloudy*
Landsat	*Iraq*	*20/12/2022*	*One large flaring smoke*

Table 5.1: Overview of the selected single observations for Sentinel-2 and Landsat-8/9.

For GOES, the intercomparison is done on the two case studies presented in Groshenry et al. 2024:

- An emission in Mexico (-104.3308, 26.0941) on 12th May 2019.
- An emission in Idaho (-116.6542, 43.7443) on 12th October 2023.

5.2. Intercomparison results for single observations

For the Landsat-8/9 and Sentinel-2 single-image intercomparison, the challenging observational conditions of the two scenes were anticipated to result in significant discrepancies between the two retrieval methods but also reduced signal-to-noise ratios. Over the 8 scenes, only one plume was detected, from UPV only, on the Sentinel-2 Iraq scene acquired on 2022, January 25th. Kayrros did not report any plume for this scene. Visual analysis of Figure 5.2 confirms the absence of discernible plume patterns or strong enhancements in the Kayrros L2 product where UPV delineated the plume.

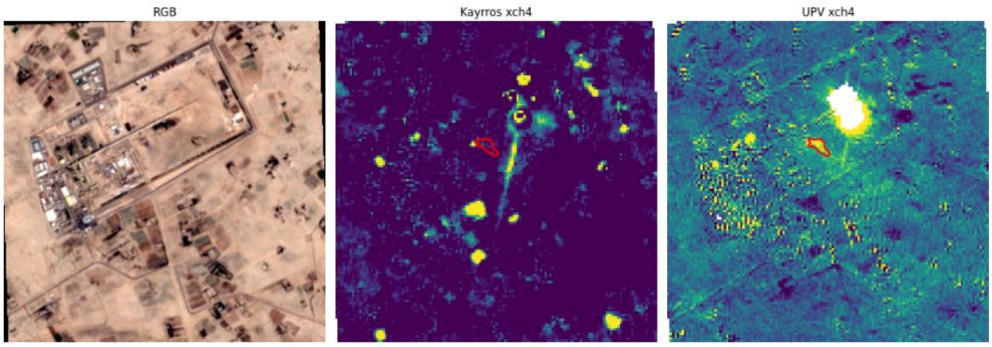


Figure 5.2: Comparison of L2 products (in ppm) between Kayrros (middle) and UPV (right) for the Sentinel-2 Iraq scene on 25/01/2022. The UPV delineation is displayed in red on the two images. The left image corresponds to the RGB composite image.

Visual inspection of the UPV and Kayrros L2 products reveals several qualitative differences in retrieval characteristics:

- **Cloud handling**: Clouds present significant challenges for both algorithms. In Kayrros retrievals, clouds are not masked but exhibit values at or near zero. Conversely, UPV applies cloud masking: however, imperfect mask boundaries result in artificially intense values at cloud edges in several scenes.

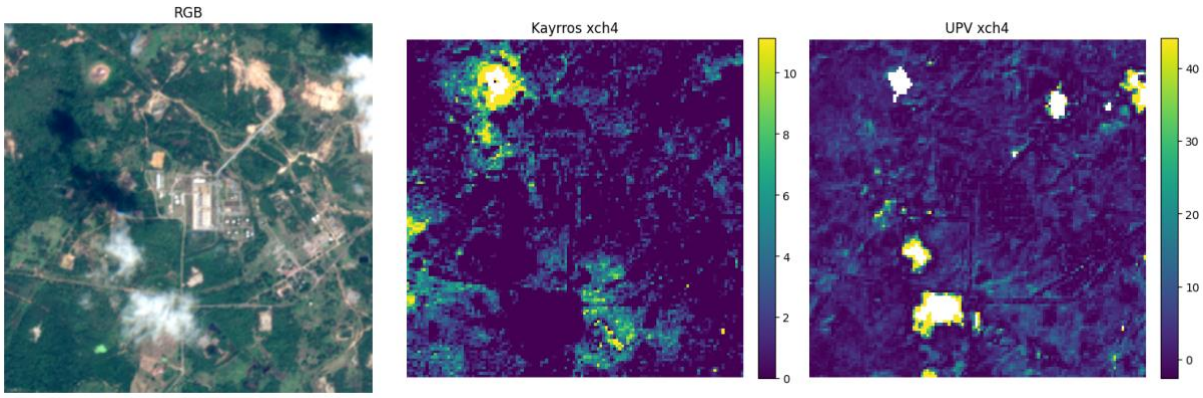


Figure 5.3: Comparison of L2 products (in ppm) between Kayrros (middle) and UPV (right) for the Sentinel-2 Venezuela scene on 22/11/2023. The left image corresponds to the RGB composite image.

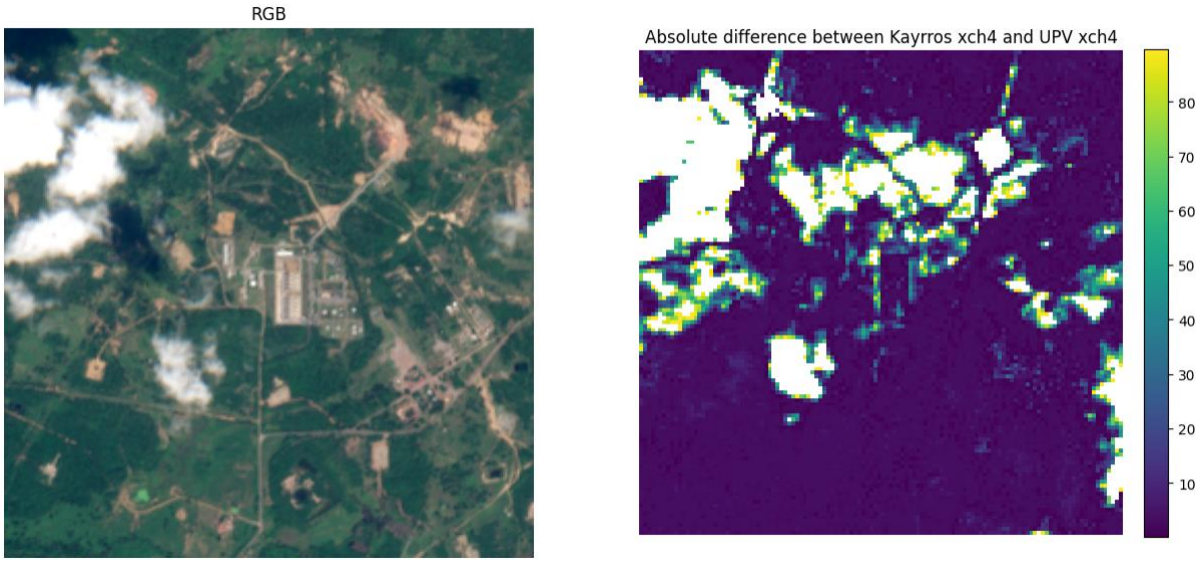


Figure 5.4: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Venezuela scene on 22/11/2023. The left image corresponds to the RGB composite image.

- **Flare and smoke artefacts**: Gas flares produce contrasting signatures between the two methods. UPV retrievals show flares as high-intensity features, while Kayrros retrievals exhibit suppressed flare signals.

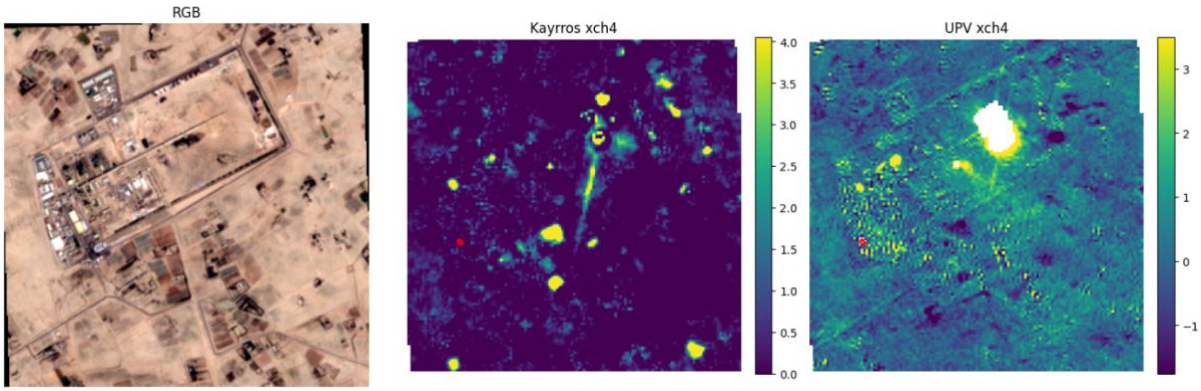


Figure 5.5: Comparison of L2 products (in ppm) between Kayrros and UPV or the Sentinel-2 Iraq scene on 25/01/2022. The left image corresponds to the RGB composite image.

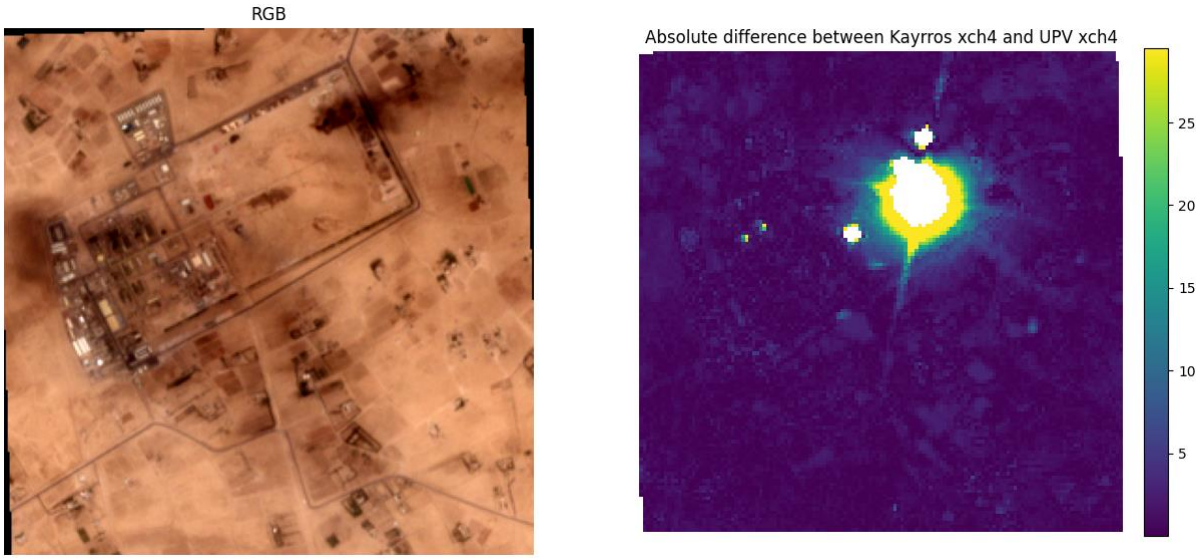


Figure 5.6: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Iraq scene on 22/11/2023. The left image corresponds to the RGB composite image.

- **Urban areas**: Spatial regions containing buildings typically exhibit reduced clarity in both retrieval products.

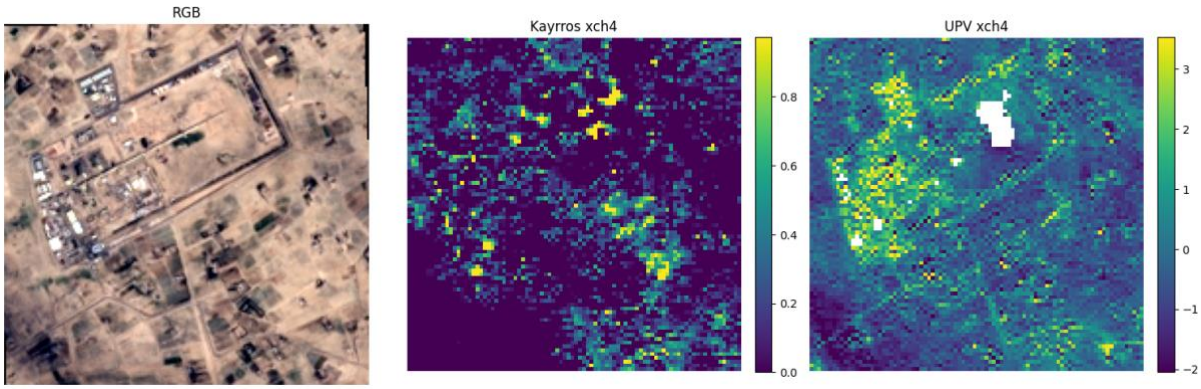


Figure 5.7: Comparison of L2 products (in ppm) between Kayrros (middle) and UPV (right) for the Landsat-9 Iraq scene on 20/12/2022. The left image corresponds to the RGB composite image.

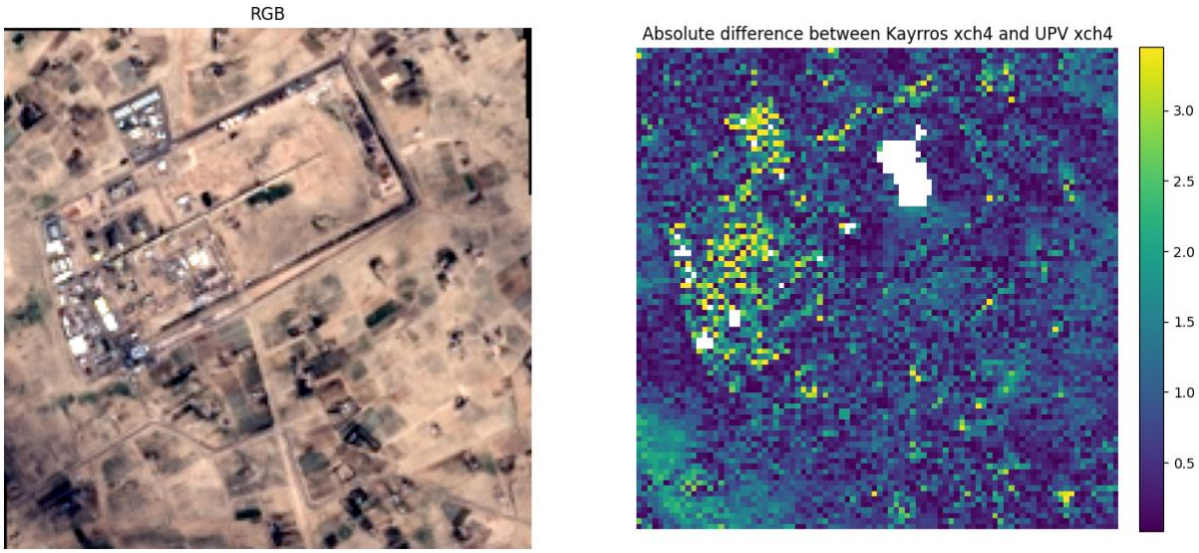


Figure 5.8: Absolute difference (in ppm) between Kayrros and UPV L2 products for the Sentinel-2 Iraq scene on 20/12/2022. The left image corresponds to the RGB composite image.

Overall, the L2 products demonstrate two primary differences: Kayrros retrievals exhibit higher noise levels with lower methane enhancement magnitudes compared to UPV products. This increased noise is likely attributable to the automated background selection methodology, where automatic selection of ten historical reference images may introduce unexpected artifacts into the retrieval process.

This observation is corroborated by quantitative analysis. Figures 5.9 and 5.10 present evaluation metrics including Extreme_MAE and Extreme_RMSE, which correspond to the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) calculated specifically for methane concentration values exceeding the 95th percentile threshold. The results demonstrate substantial discrepancies between Kayrros and UPV

methane enhancement, no matter the site or the constellation. These differences become more pronounced at extreme values - a critical consideration given that such high-concentration regions are of primary interest for methane plume detection applications.

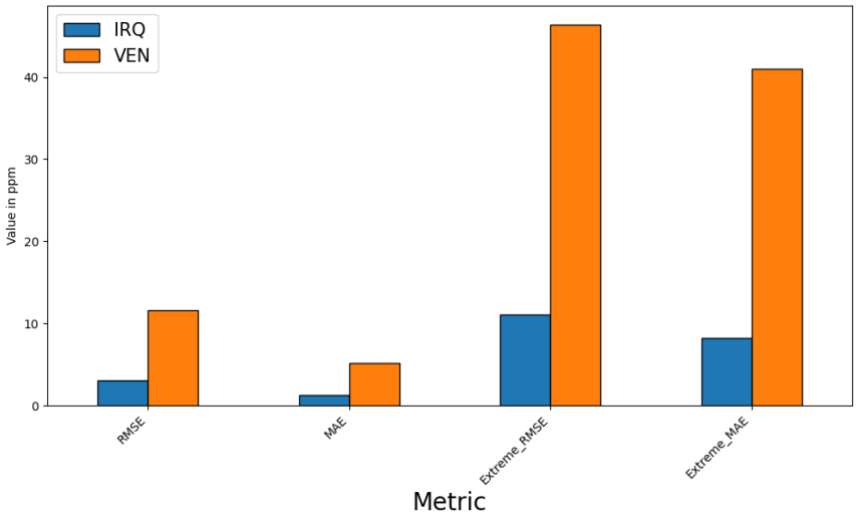


Figure 5.9: Comparison metrics averaged (in ppm) by site for the eight L2 products

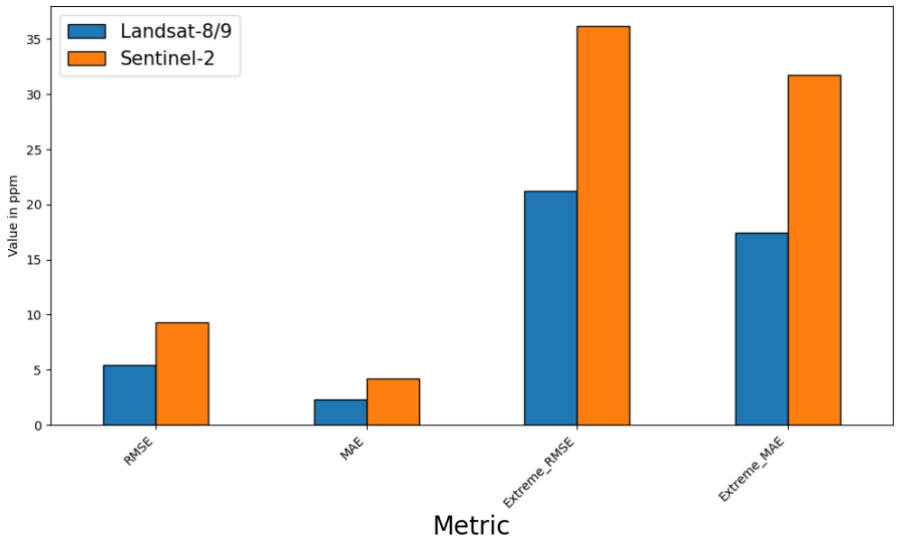


Figure 5.10: Comparison metrics averaged (in ppm) by constellation for the eight L2 products

Having established that substantial differences exist between L2 products, it would be valuable to determine whether these discrepancies primarily originate from the temporal normalization methodology or from the conversion process that transforms residual or transmittance images into concentration enhancements. Such an analysis would require more detailed and intermediary information from the respective algorithm providers, to make sure we compare apple to apple.

The GOES intercomparison examines two case studies from published literature (Watine-Guiu et al. 2023, and Groshenry et al. 2024). Figure 5.11 demonstrates that for these events, the relative difference between UPV and Kayrros estimates is approximately 40%, consistent with the magnitude observed for Sentinel-2 and Landsat-8/9. The main difference is that the uncertainty intervals for the GOES quantification results do not overlap. This result stems from the significantly lower relative uncertainties achieved by both providers: UPV reported 5% and 6%, while Kayrros reported 3% and 6%. These low relative uncertainty values are due to the flow rate calculation methodology for geostationary satellites, which is performed without relying on any external wind data.

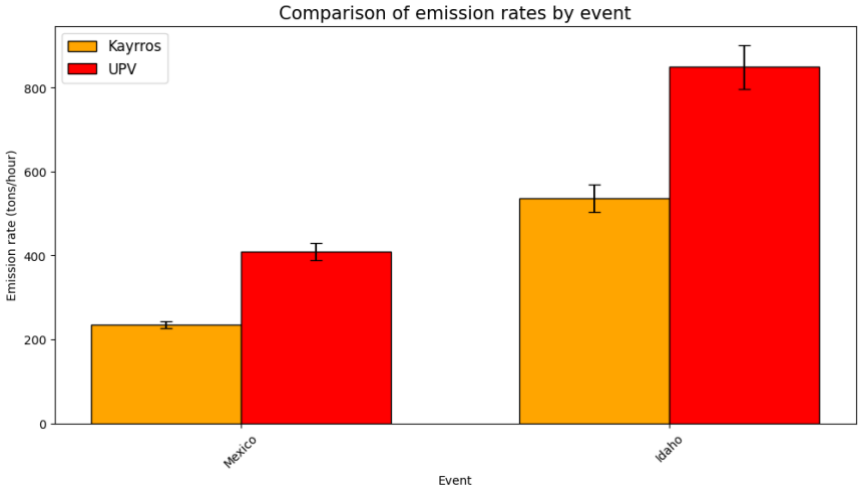


Figure 5.11: Comparison of emission rates for GOES events.

Understanding this difference requires considering how emission rates are derived from geostationary satellites. Unlike high-resolution satellites that provide single snapshots, GOES captures temporal sequences of release events with 5-minute intervals for these two cases. Emission rates are calculated from time-series analysis of Integrated Mass Enhancement (IME). The subsequent analysis will therefore concentrate on examining these time-series datasets.

As explained in Zhou et al. 2025, a release event can be segmented in different phases with:

1. The **initial emitting stage**. The plume remains compact and close to the source.
2. The **stabilization and accumulation phase**. The plume detaches from the source and evolves under atmospheric transport.
3. The **dispersion phase**. The methane disperses and mixes with the atmosphere vertically and horizontally.

Depending on the observing conditions and the timestamp of the release, it is sometimes possible to clearly distinguish the three phases. For example, as shown by

Figure 5.12, in the Idaho case, based on the time-series provided by UPV, we could say that phase 1 is between 16:40 and ~18:00, then phase 2 between ~18:00 and 19:30 and finally phase 3 after 19:30. The phase distinction is much more difficult with the time-series provided by Kayrros. For the Mexico event, GOES observations appear to have captured an incomplete temporal record of the methane emission, with solar illumination conditions ending before emission cessation. This truncated observation period explains why the IME time-series cannot be segmented into three distinct phases.

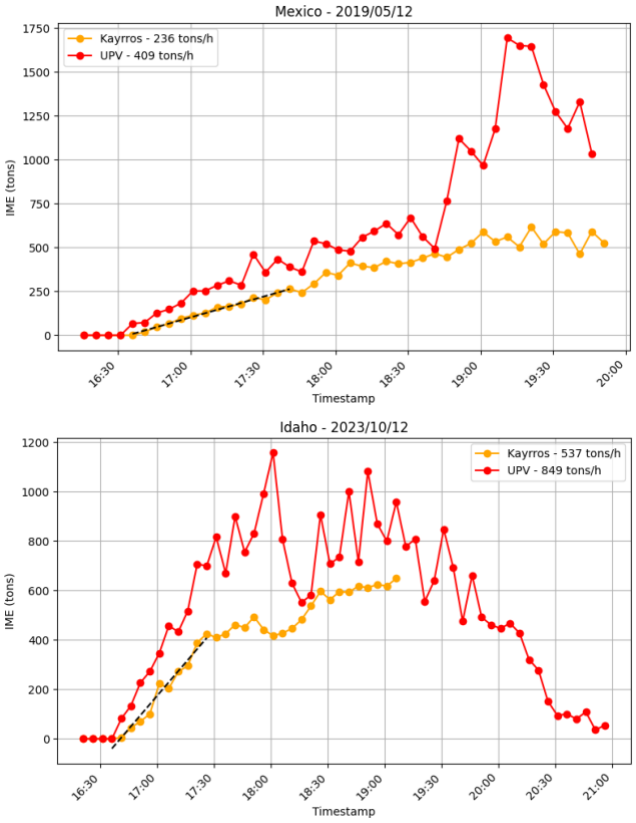


Figure 5.12: Comparison of IME time-series for GOES events (Mexico and Idaho).

Figure 5.12 clearly demonstrates that Kayrros consistently produces lower IME estimates compared to UPV. The relative differences are more pronounced during periods when the UPV time-series exhibits spikes, such as around 18:00 for the Idaho event and between 19:00-20:00 for the Mexico event.

For geostationary satellite observations, a key advantage lies in the flow rate quantification methodology, which eliminates dependency on wind data—typically a major source of uncertainty in emission estimates. The quantification relies exclusively on IME time-series analysis. Flow rate calculation is achieved through linear regression analysis of the IME time-series during the emission phase, under the assumption of

linear flow rate growth throughout this period. The linear regression fitting is illustrated in the Kayrros time-series presented in Figure 5.12.

Given this methodology, the observed discrepancies between Kayrros and UPV quantifications can come from variations in methane concentration maps (L2 products), and/or differences in the plume delineation. Both parameters were subsequently investigated to identify the source of quantification variance.

To assess the impact of masking methodology on emission quantification, an isolation analysis was conducted comparing IME time-series derived from identical L2 products (either UPV or Kayrros retrievals) applied with different masking approaches. Figure 5.13 presents the temporal results for all four possible combinations across each emission event.

The analysis reveals minimal variance in IME time-series when using a consistent L2 product with different masks. The masking approach demonstrates negligible influence on the quantification results, with one notable exception: the Idaho event when processed with UPV retrieval data, which exhibits significant deviation between the two masking methodologies.

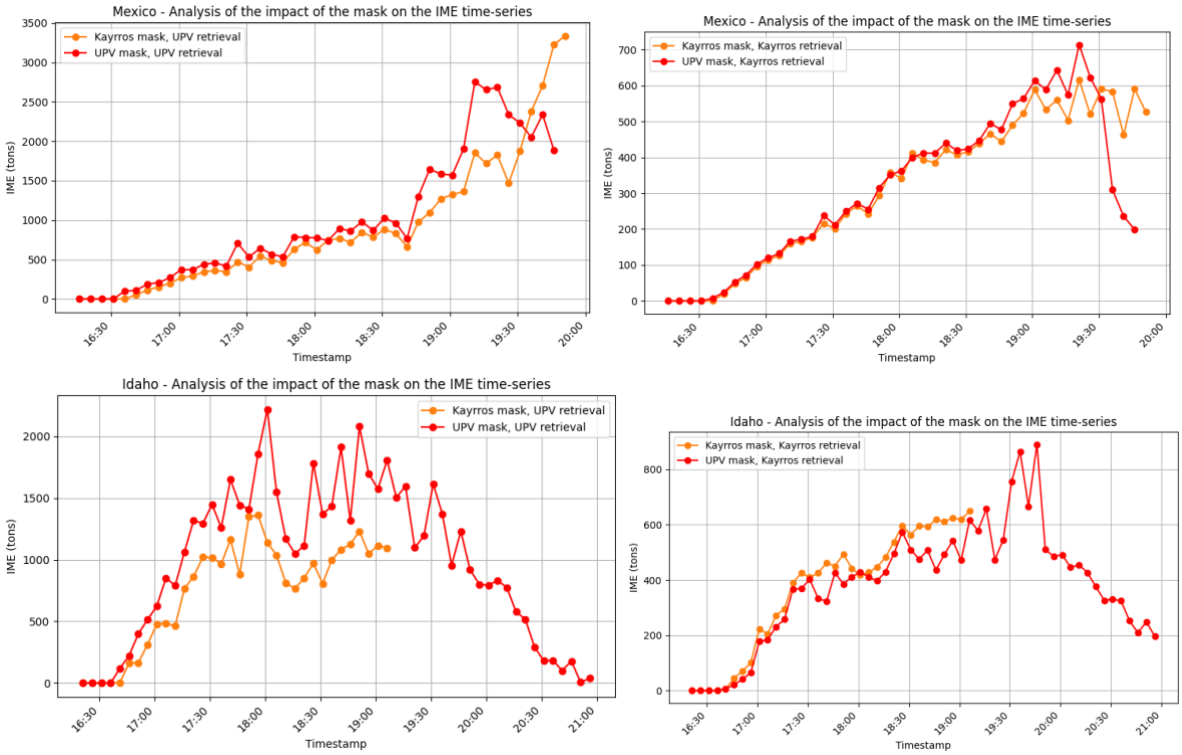


Figure 5.13: Comparison of IME time-series for GOES events, with a common retrieval but different masks.

A detailed examination of the masking methodology for the Idaho case using UPV retrieval reveals significant masking discrepancies. If applied to the UPV retrieval,

the Kayrros mask would demonstrate excessive restrictiveness , resulting in a high rate of false negatives. This masking incompatibility is clearly demonstrated in Figure 5.14, which provides a temporal snapshot illustrating the substantial under-detection of the emission plume when the Kayrros mask is applied to UPV retrieval data.

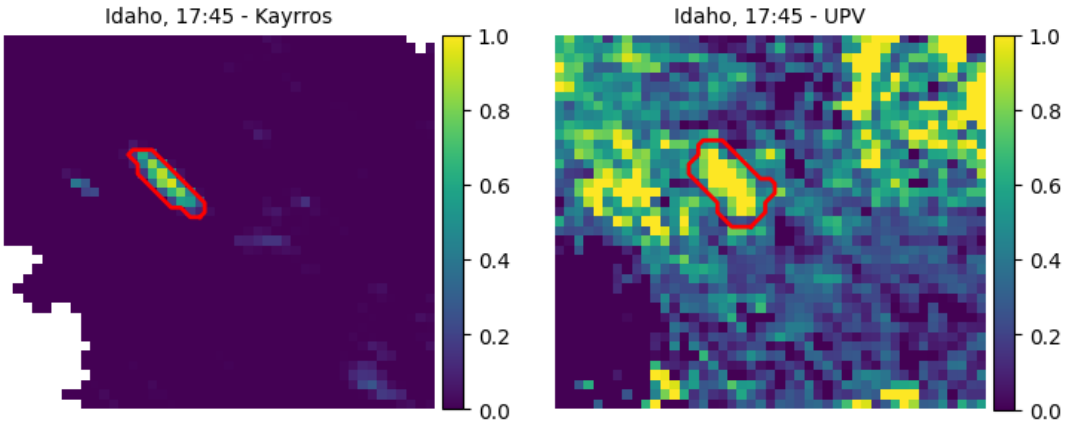


Figure 5.14: Comparison between Kayrros and UPV retrievals and masks at 17:45 UTC for the Idaho event. On the left, Kayrros methane retrieval (in ppm) with the Kayrros mask delineated in red. On the right, UPV methane retrieval (in ppm) with the UPV mask delineated in red.

To evaluate the influence of methane retrieval algorithms on emission quantification, a similar isolation analysis was performed comparing IME time-series generated using identical masks (either UPV or Kayrros masks) with different methane retrieval products. Figure 5.15 displays the temporal results across all four possible combinations for each emission event.

The analysis reveals substantial discrepancies between retrieval methods, clearly identifying methane retrieval algorithms as the primary source of flow rate differences between UPV and Kayrros quantifications. Figure 5.16 demonstrates that methane retrieval distributions within the mask intersection areas exhibit significant statistical differences. Notably, the first quartile of UPV quantification values exceeds the third quartile of Kayrros values in both cases, indicating systematic bias between the two retrieval methodologies.

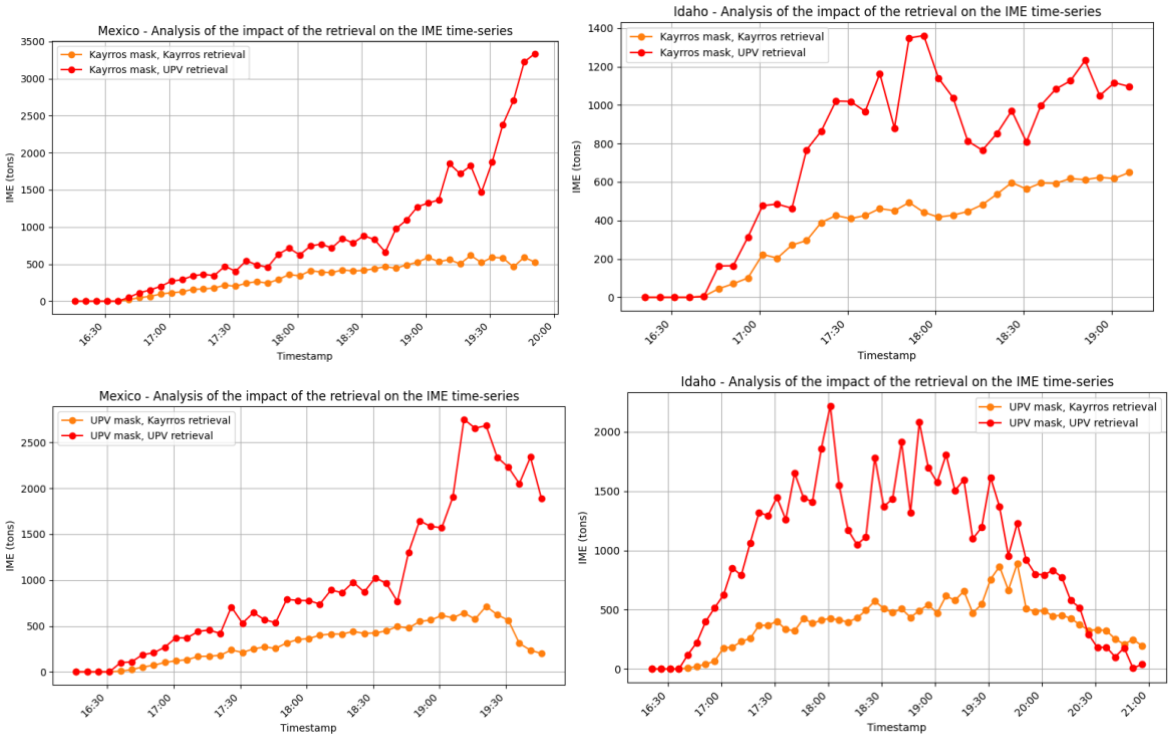


Figure 5.15: Comparison of IME time-series for GOES events, with a common mask but different methane retrieval.

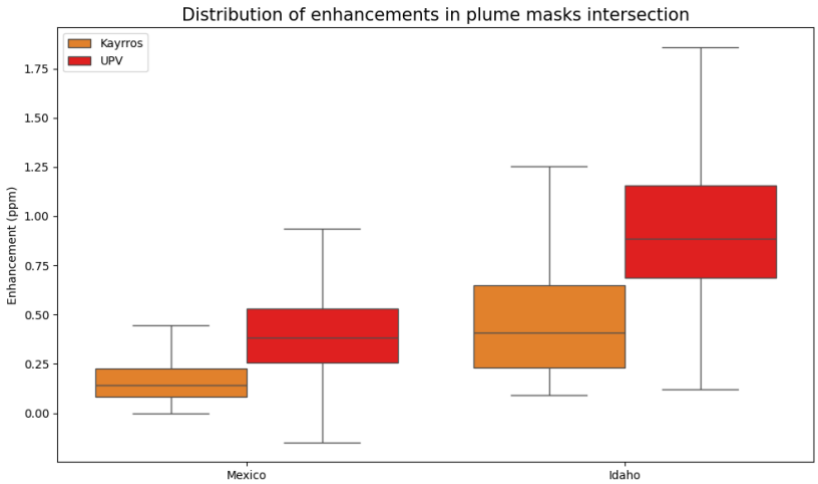


Figure 5.16: Boxplots of methane retrieval in the intersection of the two masks, for each event

To conduct a statistical analysis of this systematic bias, affine and linear regression fits were applied to pixel values within the mask intersection areas for both emission events, as presented in Figure 5.17. However, the resulting R^2 values are insufficient to support statistically significant interpretations of the relationship between the two retrieval methodologies.

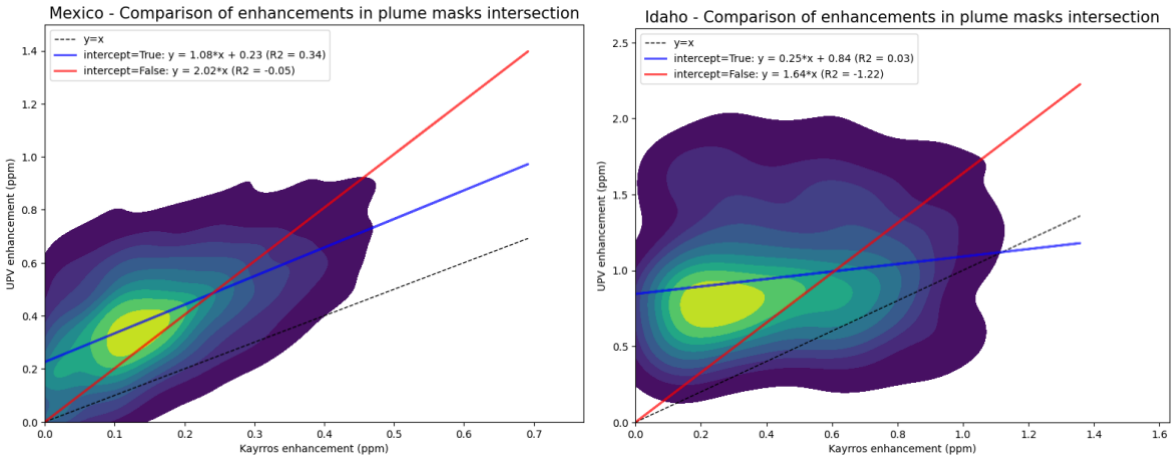


Figure 5.17: Density plots with linear and affine fits of methane retrieval in the intersection of the two masks, for Mexico (left) and Idaho (right)

Finally, retrieval methodology differences are evident in background pixel estimation, where Kayrros exhibits predominantly zero-valued pixels. While Kayrros demonstrates lower plume intensity compared to UPV, it produces fewer artifacts and reduced high-value anomalies outside designated mask areas. It is also noteworthy that cloud pixels flagged as NaN in the Kayrros retrieval correspond to regions where UPV retrievals exhibit relatively high methane concentration values. Figure 5.18 and Figure 5.19 clearly illustrate these major differences in the two retrievals.

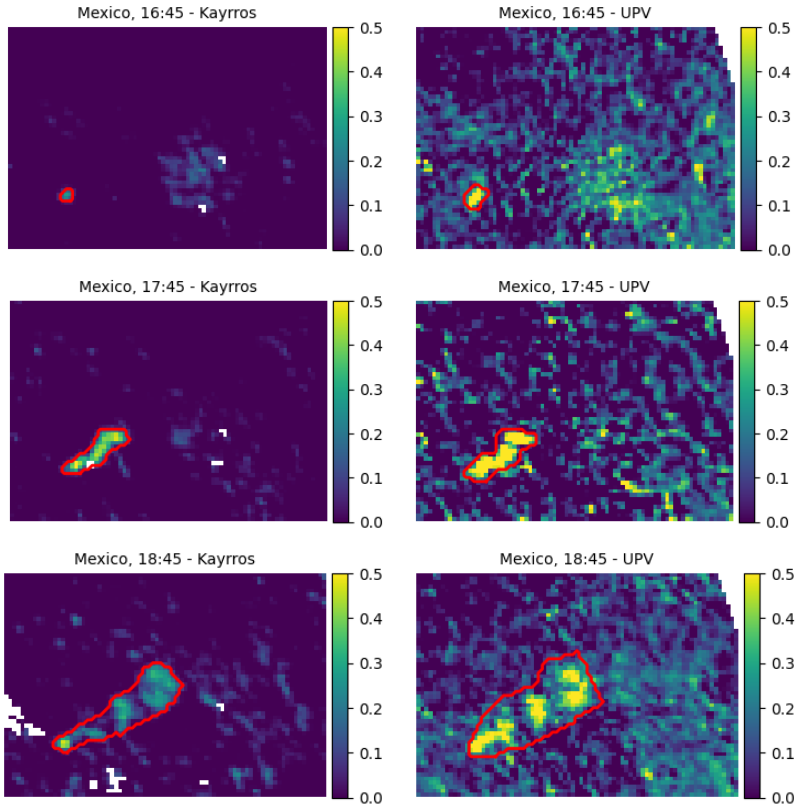


Figure 5.18: Comparison of methane retrieval (in ppm) at three different timestamps for the Mexico event. On the left, Kayrros methane retrieval (in ppm) with the Kayrros mask delineated in red. On the right, UPV methane retrieval (in ppm) with the UPV mask delineated in red.

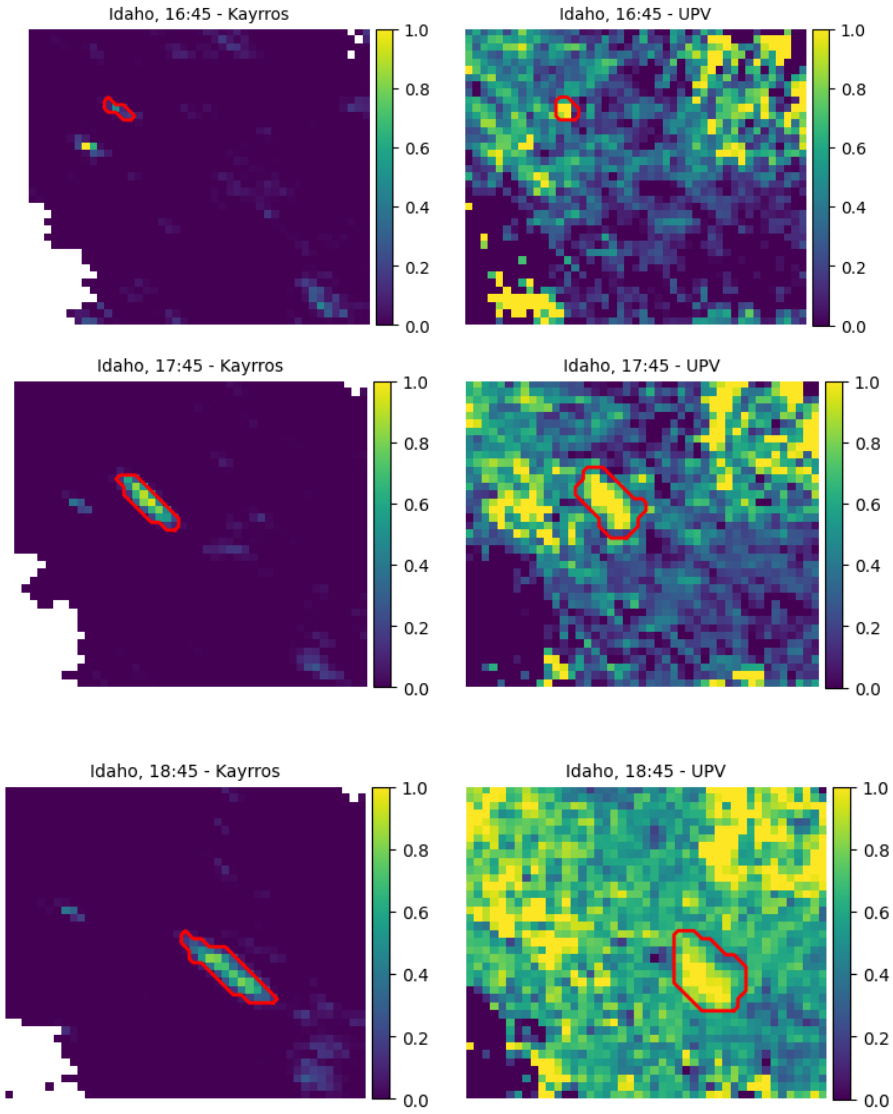


Figure 5.19: Comparison of methane retrieval at three different timestamps for the Idaho event. The mask is delineated in red.

5.3. Summary of comparisons for single observations

Comparison of GOES quantification (L4 products)

- Substantial differences in flow rates estimates between the two providers were observed for GOES use-cases, with relative differences of approximately 40%, exceeding reported uncertainty bounds.
- The primary source of these differences originates from variations in the methane L2 products, while mask delineations exhibit negligible impact on the final quantification results.
- Geostationary satellites offer the advantage of flow rate quantification without external data dependencies, eliminating wind data requirements that typically represent the dominant error component in emission rate estimates. Flow rates can be derived directly from Integrated Mass Enhancement (IME) time-series analysis. However, achieving precise L4 products necessitates improved uncertainty characterization of the XCH4 L2 retrievals.

Comparison of Sentinel-2 and Landsat-8/9 methane enhancement (L2 products)

- As expected, given the low spectral resolution of multispectral satellites, substantial discrepancies were observed between UPV and Kayrros, for Landsat-8/9 and Sentinel-2 L2 products, primarily attributed to challenging observational conditions in the Iraq and Venezuela sites. These conditions introduced artifacts from gas flares, cloud contamination, and urban infrastructure, highlighting the critical importance of robust preprocessing methodologies and artifact removal procedures.
- Future work should focus on identifying whether observed differences originate from background removal techniques or methane retrieval algorithms applied to residual or reflectance ratio data. This will be accomplished through systematic step-by-step algorithm comparison using identical input datasets to isolate the source of discrepancies at each processing stage.

6. References

Buchwitz, M., Roger, J., Groshenry, A., Schneising-Weigel, O., Hilker, M., Noël, S., Reuter, M., Hachmeister, J., Boesch, H., Dogniaux, M., Zhang, X., Maasakkers, J. D., Aben, I., Gorroño, J., Guanter, L., Algorithm Intercomparison Plan (AIP), WP310 D3.1, version 1.1, Doc ID: SRON-ESG-PL-2024-003, Date: 16-12-2024, pp. 59, 2024.

Duren, R.M., Thorpe, A.K., Foster, K.T. et al.: California's methane super-emitters. Nature 575, 180–184 (2019). https://doi.org/10.1038/s41586-019-1720-3

Ehret T., De Truchis A., Mazzolini M., Morel J.-M. and Facciolo G.: Automatic Methane Plume Quantification Using Sentinel-2 Time Series, IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 1955-1958, doi: 10.1109/IGARSS46834.2022.9884134.

Frankenberg, C., Thorpe, A.K., Thompson, D.R., Hulley G., Kort, E., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., Conley, S., Bue, BD., Aubrey, AD., Hook, S., Green R.O.: Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region. Proc Natl Acad Sci U S A. 2016 Aug 30;113(35):9734-9. doi: 10.1073/pnas.1605617113. Epub 2016 Aug 15. PMID: 27528660; PMCID: PMC5024631.

Groshenry, A., Giron, C., Hessel, C., de Franchis, C., Facciolo, G., Ehret, T.: Methane emissions monitoring using geostationary satellites, IGARSS 2024, IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024.

Heidinger, A., Evan, A., Foster, M., Walther, A.: A Naive Bayesian Cloud-Detection Scheme Derived from CALIPSO and Applied within PATMOS-x, Appl. Meteorol. Climatol. 51, 1129–1144 (2012).

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez,

P., Munoz-Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, ERA Report Series, https://doi.org/10.21957/tkic6g3wm, 2018.

Jacob, D. J., Turner, A. J., Maasakkers, J. D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., and Frankenberg, C.: Satellite observations of atmospheric methane and their value for quantifying methane emissions, Atmos. Chem. Phys., 16, 14371–14396, https://doi.org/10.5194/acp-16-14371-2016, 2016.

Thompson, D. R., Thorpe, A. K., Frankenberg, C., Green, R. O., Duren, R., Guanter, L., Hollstein, A., Middleton, E., Ong, L., and Ungar, S.: Space-based remote imaging spectroscopy of the Aliso Canyon CH4 superemitter, Geophys. Res. Lett., 43, 6571–6578, https://doi.org/10.1002/2016GL069079, 2016.

Thorpe, A. K., Frankenberg, C., and Roberts, D. A.: Retrieval techniques for airborne imaging of methane concentrations using high spatial and moderate spectral resolution: application to AVIRIS, Atmos. Meas. Tech., 7, 491–506, https://doi.org/10.5194/amt-7-491-2014, 2014.

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, Atmos. Meas. Tech., 11, 5673–5686, https://doi.org/10.5194/amt-11-5673-2018, 2018.

Varon, D. J., Jervis, D., McKeever, J. and Spence, I. and Gains, D. and Jacob, D. J.: High-frequency monitoring of anomalous methane point sources with multispectral Sentinel-2 satellite observations, Atmos. Meas. Tech., 14, 2771-2785, 10.5194/amt-14-2771-2021, 2021.

Vaughan, A., Mateo-Garcia, G., Irakulis-Loitxate, I., Watine, M., Fernandez-Poblaciones, P., Turner, R. E., Requeima, J., Gorroño, J., Randles, C., Caltagirone, M. and Cifarelli, C. :AI for operational methane emitter monitoring from space, arXiv preprint arXiv:2408.04745, 2024.

Watine-Guiu, M., Varon, DJ., Irakulis-Loitxate, I., Balasus, N., Jacob, D.J.: Geostationary satellite observations of extreme and transient methane emissions from oil and gas infrastructure. Proc Natl Acad Sci U S A. 2023 Dec 26;120(52): e2310797120. doi: 10.1073/pnas.2310797120. Epub 2023 Dec 19. PMID: 38113260; PMCID: PMC10756283.

Zhou, S., Gorrono, J., Roger, J., Irakulis-Loitxate, I., Lindstrot, R., Pei, Z., Si, L., Guanter, L.: Assessing the Potential of the MTG-FCI Geostationary Mission for the Detection of Methane Plumes. doi: 10.31223/X5WB32,2025

***** End of Document *****