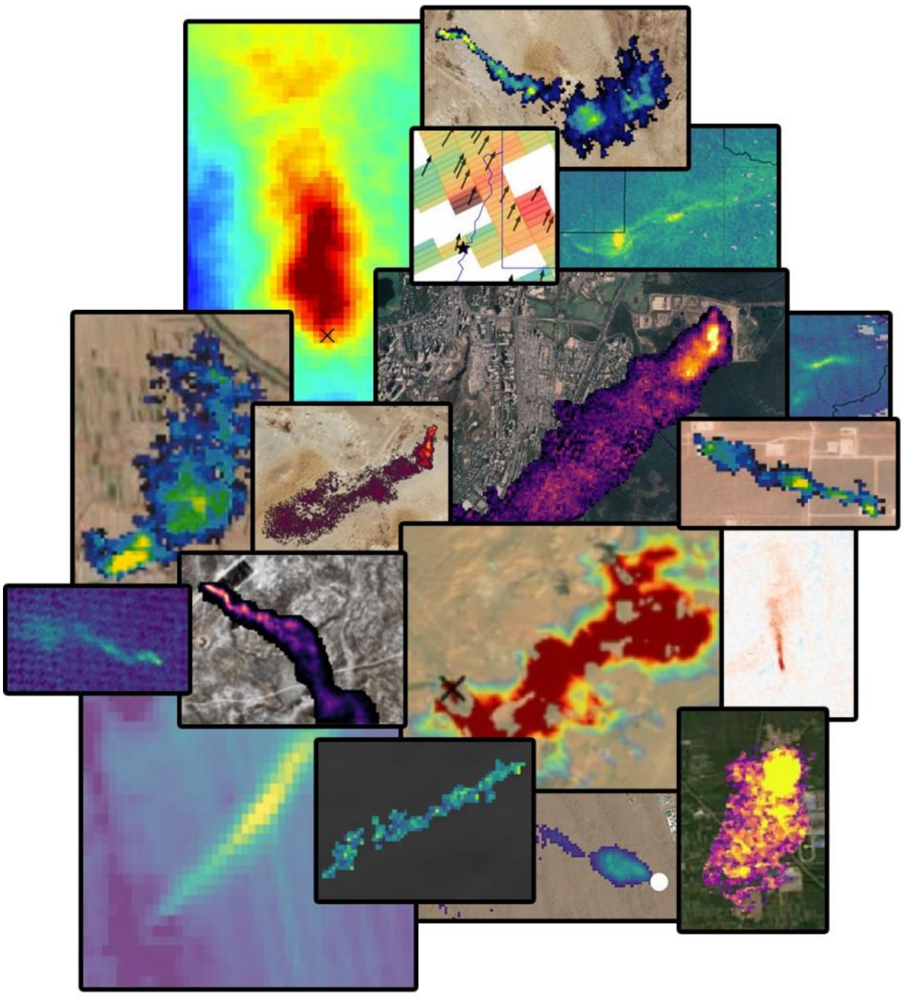
Product Validation and Intercomparison Protocol (version 2)

WP210 - Deliverable D2.1



ESA Contract No. 4000143908/24/I-LR

Authors

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

- Luis Guanter (WP210 & document lead)
- Javier Gorroño

Kayrros SAS

- Julian Akani Guery

SRON Netherlands Institute for Space Research

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

Royal Belgian Institute for Space Aeronomy (BIRA)

- Bart Dils
- Bavo Langerock
- Mahesh Kumar Sha

Institute of Environmental Physics (IUP), University of Bremen

- Michael Buchwitz
- Oliver Schneising
- Jonas Hachmeister
- Stefan Nöel

GHGSat

- Antoine Ramier

Change log

Version	Date	Status	Authors	Reason for change
0.1	30-July-2024	Draft	see Authors list	New document
1.0	31-Oct-2024	Version submitted to ESA	see Authors list	Revision after internal feedback
1.1	16-Dec-2024	Resubmitted to ESA	See Authors list	Revised based on RIDs received (Date)
2.01	16-Jan-2026	Draft	Matthieu Dogniaux	V1.0 copied to initiate efforts towards V2.0
2.03	23-Feb-2026	Draft	See Authors list & Matthieu Dogniaux	Issue fix: merging V2.01 draft with V1.1 changes
2.04	22-Apr-2026	Draft	See Authors list & Matthieu Dogniaux	Consolidating draft after many contributions
2.0	08-May-2026	Version submitted to ESA	See Authors list	Final adjustments and version number change from 2.04 -> 2.0 for submission to ESA
2.1	31-July-2026	Version submitted to ESA	See Authors list	Included feedback expert review

Approved

Date: 31-07-2026 Signature:	Date: Signature: Simon Pinnock Digitally signed by Simon Pinnock Date: 2026.08.12 09:22:24 +01'00'
Dr. Ilse Aben Science Lead, SRON	Dr. Simon Pinnock ESA Technical Officer

Table of Contents

1 Preface

This document presents a protocol for the validation and intercomparison of point-source methane flux products from space-based instruments. It corresponds to the work package (WP) 210 of the ESA *Methane Emissions Detection Using Satellites Assessment* (MEDUSA) project.

The proposed product validation and intercomparison protocol (PVIP) is split into two interrelated workflows: the first one is devoted to hyper- and multispectral instruments with a spatial resolution between 20 and 60 m, and a sufficiently low plume detection limit to allow for the comparison with ground-based controlled methane releases; the second workflow is for coarser resolution missions (mostly global flux mappers such as TROPOMI) which cannot capture single plumes from controlled release experiments, and therefore true direct validation is not possible. In this case, we assess their performance based on multiple (inter-)comparisons with data from the validated high-resolution systems.

The MEDUSA project is designed to have two subsequent product validation and intercomparison cycles within WP200. The first version of this document served as the basis for the first 12-month, run in WP220, "System Implementation and Demonstration". This second version of the document incorporates the lessons learned during that first cycle as discussed in the product validation and intercomparison report (PVIR) (Dils et al., 2026). This includes improvements in the data analysis methods, and updates to the datasets considered for the different intercomparisons. It also describes the validation and intercomparison tasks and tests to be performed during MEDUSA's second validation and intercomparison cycle.

This document is structured in four major Sections following this preface. Section 2 discusses the requirements identified for the validation and intercomparison exercise, and how these are addressed by the proposed protocol. Sections 3 and 4 describe the workflows for the validation and intercomparison of high-resolution and flux-mapping satellites, respectively. Section 5 details a number of technical aspects related to the particular implementation and application of the protocol within MEDUSA. Finally, Section 6 summarizes the main points of the validation and intercomparison protocol.

2 Background

Following ESA's Statement of Work, MEDUSA's WP210 is expected to "*develop and implement a pre-operational system able to demonstrate routine sustained intercomparison and validation on a common basis of satellite-based greenhouse gas emission hotspot information*", address the following specific points:

- Represent a wide set of areas of interest, range of emission sources and atmospheric, background and illumination conditions.
- Provide information on the quality of each product regarding hotspot identification, location, intermittency, false alarm rates, and omission errors.
- Develop and apply a product quality evaluation matrix to summarize the results of the assessment for each included product.
- Provide an inventory of reference data available for use in validation, including source and access conditions.
- Be applicable to products from the required instruments that have widely varying measurement capabilities.
- Define a common data format to facilitate the intercomparison of data sets from different instruments and providers.
- Identify any development needs to improve the systematic intercomparison and validation of hotspot emissions.
- Be implemented as a pre-operational system and demonstrated through two 12-month cycles of operation

In our interpretation of these requirements, we are requested to define a generic framework for the systematic validation and intercomparison of satellite-based methane flux rate estimates from point source emissions, referred to as "flux products" hereinafter. This protocol shall accommodate the wide range of satellites and flux products under consideration in MEDUSA.

By validation, we understand the assessment of the quality of the satellite-based flux products by comparison with a reference standard that must be independent and have an uncertainty significantly smaller than the compared quantity; by intercomparison, we refer to the comparison of a given flux product to equivalent flux products derived from different satellites or algorithms. From the validation exercise we expect to assess the compliance of flux products against a benchmark, whereas from the intercomparison exercise we can infer the potential and limitations of different data products relative to other comparable products.

We divide the validation and intercomparison framework in two specific workflows, one for high-resolution satellites and one for flux mappers or any other satellites for which flux products cannot be directly compared to controlled releases. The reason for this split is that the data from ground-based controlled methane releases, which provide the most reliable validation reference, is a useful validation reference only for the high-resolution missions, which have the sensitivity and spatial resolution required for the detection and quantification of the plumes generated during controlled release tests. For the coarser-resolution satellites, the intercomparison with data from validated high-resolution products will keep the traceability of the flux products all the way down to the controlled releases as much as possible.

An important limitation with which we have to deal is the limited number of match-ups between high-resolution satellite acquisitions and controlled-releases: first, there is no global network of controlled releases operating under various atmospheric observation conditions; second, the spatio-temporal sampling of the hyperspectral missions is typically sparse, and the sensitivity of the multispectral systems is low, which implies that only a reduced number of matchups between in situ and satellite-based flux rate values can be achieved. Also, controlled-release experiments rely on complex set-ups and heavy infrastructure that complicate the deployment at various sites so different surface and atmospheric conditions could be sampled. Finally, the intercomparison of flux rate estimates from different satellites is strongly constrained by the rare occurrence of quasi-simultaneous observations of the same plume by different satellites.

In order to alleviate the issue of the limited number of validation points and sites, we propose to use simulated plumes anchored to controlled-release data. The simulated plumes will be embedded in real top-of-atmosphere data and validated through comparison with real plumes from the controlled-releases. After that, simulated plumes will be used to extend the number and range of the validation points, and also to extend the validation to other sites for which top-of-atmosphere data are available.

This scheme linking validation and intercomparison tasks, including simulated plumes, is represented in the flow chart depicted in Fig. 2.1. In short, the core validation of the high-resolution products will be achieved through comparison with controlled release data, with the possibility of extending the number of points and sites using simulated plumes; for flux mappers and coarse-resolution instruments in general, true direct validation is not possible, but we will perform different intercomparisons to assess their performance (e.g. with the validated high-resolution products). All these comparisons come with their own complications, as finding exactly matching emission footprints is very challenging. The intercomparison of products from the different missions and algorithms is also planned in order to learn about their relative performance.

The validation and intercomparison workflows proposed for high-resolution satellites and flux mappers are described in the following Sections.

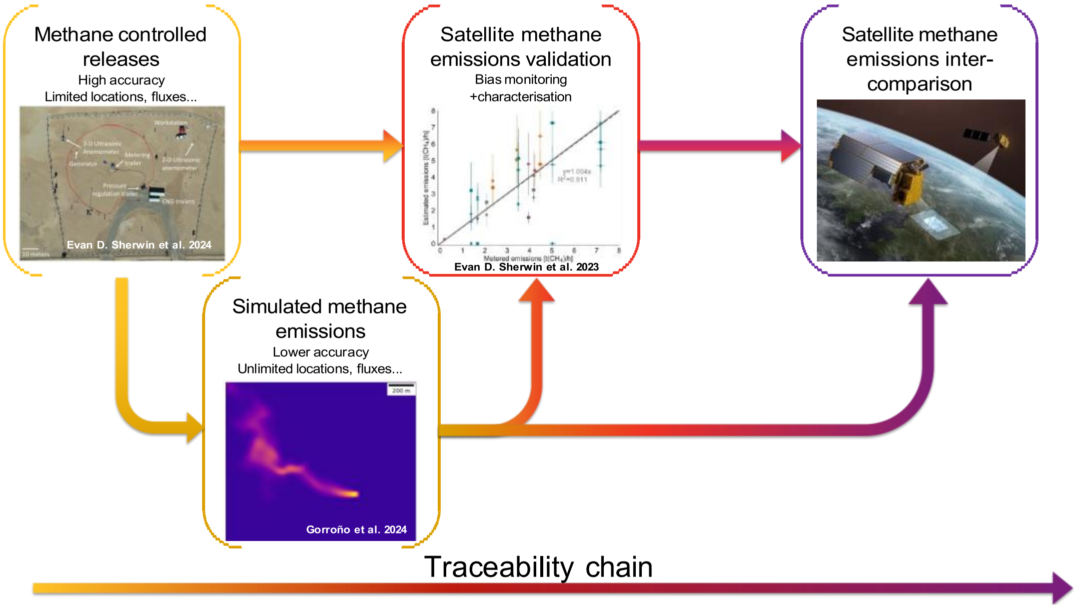


Figure 2.1: Scheme of the proposed validation and intercomparison framework. It consists of the validation of flux products from high-resolution instruments, the extension of the validation database with simulated plumes, and the intercomparison of products from different satellites and algorithms.

3 Validation and intercomparison of methane flux products from high-resolution missions

The core tasks for the validation of flux products from high-resolution instruments will rely on ground-based controlled-release data, which will be used as a validation reference. The available controlled-release data and the methods for the comparison with satellite-based flux products are detailed in the next Sections.

Concerning intercomparisons between flux products, we split those into inter-algorithm and inter-mission comparisons. For inter-algorithm comparisons (same input data, different processing chain), controlled release data can also be useful for an informed intercomparison of flux products: the outputs from different algorithms applied to the same satellite acquisitions over controlled releases are intercompared, and the results from that inter-comparison are interpreted with the support of the controlled release data. Finally, flux products from quasi-simultaneous observations by different satellites over stable sources will also be intercompared.

The controlled release data points and sites used for validation will be extended with simulated plumes, which we consider scientific tasks (i.e. based on ad hoc tests and analysis, rather than the more operational nature of the core tasks). These simulations will also enable the intercomparison of flux rate estimates from different satellites by means of embedded identical simulated plumes in the acquisitions by the different satellites.

The core validation and intercomparison tests as well as the extension of those using simulated plumes, are summarised in Fig. 3.1. The details of those tests are discussed in the remainder of this Section.

3.1 Existing and planned controlled release datasets

Three main controlled-release experiments including public satellite data were carried out in 2021, 2022 and 2024-2025 by a team led by Adam Brandt at Stanford University. These experiments (referred to as “Stanford campaigns” hereinafter) enabled single-blind testing of flux rate estimates from a number of high-resolution satellites. The three tests differed in the satellites involved, the location, and the range of released methane volumes, as follows:

- **Stanford campaign 2021 (Sherwin et al., 2023):** The first campaign took place at a desert site in Ehrenberg, Arizona (USA) over a 19-day period from 16 October to 3 November, 2021. Natural gas with a high concentration of methane was emitted through release stacks at a height of either 7.3 or 3 m above ground level. The metered flux rates ranged from 0.2 to 7.2 metric tons per hour (t/h). Five satellites were tested (GHGSat-C2, PRISMA, Sentinel-2, Landsat 8, and WorldView-3). A total of 19 match-ups between satellite-based and metered flux rates were obtained.

Validation and intercomparison of flux products from high-resolution satellites

Core tasks

Validation of flux rate estimates through comparison with controlled releases

- Direct comparison of satellite-based flux rates with controlled release data
- Metrics on bias, correlation and other metrics to be obtained

Intercomparison of flux products from different algorithms

- Multiple data processing schemes applied to the same set of satellite observations
- It will include time series over controlled release tests to support the interpretation of the results

Intercomparison of flux products from different satellites

- Identified pairs of quasi-simultaneous observations over stable sources

Scientific tasks

Extended validation with a wider range of flux rates

- Simulated plumes to extend the range of flux rates of the controlled release dataset
- Plume simulations first validated with real controlled release plumes

Extension of the validation to different sites

- Simulated plumes embedded in radiance datasets different from the controlled release ones

Intercomparison of flux products from different satellites (with plume simulations)

- Aided by plume simulations to provide a common reference (simulated plumes embedded on temporally-close acquisitions by different satellites)

Figure 3.1: Summary of tasks for the validation and intercomparison of flux products from high resolution satellites, including *core tasks*, mostly relying on direct comparisons with collocated controlled releases, and *scientific tasks*, involving plume simulations.

- **Stanford campaign 2022 (Sherwin et al., 2024)**: The second Stanford campaign was deployed near Casa Grande, Arizona (USA). It extended the number of satellites and validation match-ups with respect to the 2021 campaign, although with a lower range of emissions. A total of 82 controlled methane releases were conducted between 10 October and 30 November 2022 during overpasses of nine satellite missions and constellations, which included the ones participating in the 2021same campaign and also the EnMAP, Gaofen-5 and Ziyuan-1 hyperspectral missions. The metered flux rates ranged from 0.3 to 1.6 t/h.
- **Stanford campaign 2024-2025 (Reuland et al., 2026)**: The last controlled release campaign by the Stanford group was the most ambitious one. It consisted of 5 quarterly phases between August 2024 and December 2025 and covered semi-arid sites in the USA, including one in Wyoming and one in Arizona. The campaign involved 15 analysis teams and tested 12 satellite systems with around 500 releases ranging between 50 and 1200 kg/h. A number of matchups with GHGSat, EnMAP, PRISMA, EMIT, Gaofen-5, Tanager-1, Sentinel-2, and Landsat acquisitions were obtained.

The three Stanford controlled-release campaigns described earlier represent the most complete source of validation data for satellite-based methane point source imagers. In addition, the Methane Emissions Technology Evaluation Center (METEC) led by Dan Zimmerle at Colorado State University promises to become an important data source for the validation of methane flux products with controlled release data. METEC is a large controlled-release facility being implemented in Colorado (USA). This facility is intended to mimic the infrastructure that one can find at an oil and gas extraction basin. Data from METEC for the validation of satellite flux products are not yet available to the best of our knowledge.

3.2 Methodology for the validation of flux products with controlled-release data

The goal is to validate satellite-based flux products through the comparison with collocated metered fluxes from controlled release experiments. However, the small number of metered vs satellite match-ups available for some comparisons is one of the main challenges in the validation of emission rates. In order to reach the most robust validation results in the event of a limited set of match-ups, we combine two statistical analysis frameworks that offer similar yet conceptionally different approaches that we will name as orthogonal regression slope and mean relative bias.

The orthogonal regression (OR) slope approach for the validation of flux products using controlled-release data builds on the one developed for the analysis of the data from the past Stanford campaigns (Sherwin et al., 2023, 2024). The basis of this methodology relies on linear regression fits as the metric for the controlled-release evaluation of sensing technologies. The fits assume an orthogonal distance regression, using the inverse of the relative emission flux uncertainties squared on both the reference and satellite-based flux products as weights, with a linear model in which the intercept is forced to 0,

$$y = \beta x \tag{3.1}$$

Note that in our initial validation results reported in MEDUSA's PVIR (Dils et al. 2026), we used the absolute uncertainty values as weights. Upon evaluation and the observation that absolute uncertainties tend to be proportional to the absolute obtained flux rates, we advise against it. The metered emission rate (average from the 5-min preceding the satellite overpass) is plotted on the x-axis, and the flux rates from the satellite product under validation are used for the y-axis. The implications of forcing the intercept to 0 are discussed in Sherwin et al. (2023). If the slope of the fitted linear model is close to the 1:1 agreement, the product is expected to be unbiased with respect to the controlled release data. The coefficient of determination of the fit (r^2) is calculated to indicate the extent to which the linear fit accurately describes variation in the data.

For each satellite flux product entering the validation exercise, we require that there is at least one matchup between cloud-free satellite acquisitions and detectable controlled releases, although at least 3 points are required for the statistical analysis. In this case, the slope β and the squared correlation r^2 of the linear fit are calculated and stored in order to allow the latter intercomparison of results. Statistical-significance variables will also be calculated for each test.

The second approach, mean relative bias (MRB), considers each collocated satellite-reference pair as an indicator of a relative bias between the two products. The relative offset between x and y thus becomes

$$\delta_{rel} = (y - x)/x \tag{3.2}$$

The overall relative offset Λ_{rel} between two paired datasets then corresponds with the weighted average of all individual δ_{rel} . Weights are based on the conservative error which stems from the combined relative uncertainty of x and y, assuming multiplicative errors.

$$w_i = \frac{1}{\sigma_{cons,i}^2} \text{ and } \sigma_{cons,i} = \sqrt{\left(\frac{\sigma_{x_i}}{x_i}\right)^2 + \left(\frac{\sigma_{y_i}}{y_i}\right)^2} \tag{3.3}$$

With σ_{xi} and σ_{yi} the respective absolute uncertainties on the retrieved flux rates x_i and y_i the satellite and reference flux products respectively.

Starting from the individual relative biases δ_{rel} one can then derive a histogram of the bias distribution and establish the 90% confidence interval of the overall relative offset Λ_{rel} using an error weighted bootstrapping approach. These histograms allow us to evaluate the normality of the distribution. To have a further indication on the statistical strength of any observed overall relative bias, we perform a Wilcoxon signed-rank and Student t-test to calculate the probability that the observed bias is significantly different from zero, under the null hypothesis (H_0) that the true bias equals zero. The difference between the two is that Wilcoxon does not assume normality of the distribution but is less robust when handling very small datasets. Additionally, a χ^2 test assesses whether the stated 1σ uncertainties adequately represent the observed scatter, by comparing the sum of squared residuals, normalized by the reported uncertainties, to a chisquare distribution with *N – 1* degrees of freedom. Note that to accurately portray the reported uncertainties one needs to multiply the conservative error with a factor f, with f being *y/x.* Note that this test assumes that the errors are uncorrelated, which may not be correct. Therefore, one should regard the χ^2 analysis of the uncertainty as indicative only.

In summary, both methods do assume that there is no measurable fixed bias between the two datasets, and that the relationship can therefore be described as a simple ratio between the satellite flux product and the reference dataset. If the comparison with reference data should nevertheless point to the presence of such an offset, the approaches outlined here should be reevaluated. In that case, an ortogonal regression slope+intercept fit with relative weights is advised to be used as the primary statistical parameter, with the relative bias assessment in a supporting role. Ignoring an intercept, orthogonal regression and mean relative bias methods produce similar but slightly different numbers because they answer slightly different questions. The orthogonal regression slope estimates the relative sensitivity of two instruments (i.e. how the datasets scale with changing signal magnitude), while the mean relative bias quantifies the typical fractional difference between paired observations. The two metrics become nearly identical when the relationship between datasets is close to proportional and the relative differences are small. Differences become more noticeable when the scatter is large, when the true relationship deviates from the assumed proportional relationship (e.g. due to a hidden intercept or non-linearity), or when the dataset spans a very large dynamic range that is unevenly sampled.

A mock-up validation exercise of three different satellite-based flux rate products compared with collocated controlled release data using the orthogonal regression slope approach is illustrated in Fig. 3.2. In this simulated example, Product A shows a high correspondence with the reference controlled release data, with a low bias (slight overestimation of 2%) and a strong correlation (Fig. 3.2a); product B represents a case with a higher positive bias than A (9%) and a weaker correlation with the metered data (r^2 decreases from 0.99 to 0.71) (Fig. 3.2b); finally, product C is strongly correlated with the controlled release data (r^2=0.90), but it is low-biased by 27% (Fig. 3.2c). The root mean squared error (RMSE) between the satellite and controlled release datasets is relatively similar for the B and C products (269 and 306 kg/h, respectively), despite the different bias and correlation, and much smaller for product A (42 kg/h) The panel in Fig. 3.2d corresponds to the comparison of Product A after extending the validation dataset with simulated plumes (see Section 3.4). Figure 3.3 shows a similar mock-up validation exercise but now following the relative mean bias approach. A again corresponds with a low-bias, high-correlation dataset, B with a higher bias, lower correlation dataset and C with a higher bias, higher correlation dataset.

In addition to the validation of flux rate estimates, for each satellite dataset used to calculate flux rate values entering the validation, it is requested to provide the raw methane (enhancement) concentration

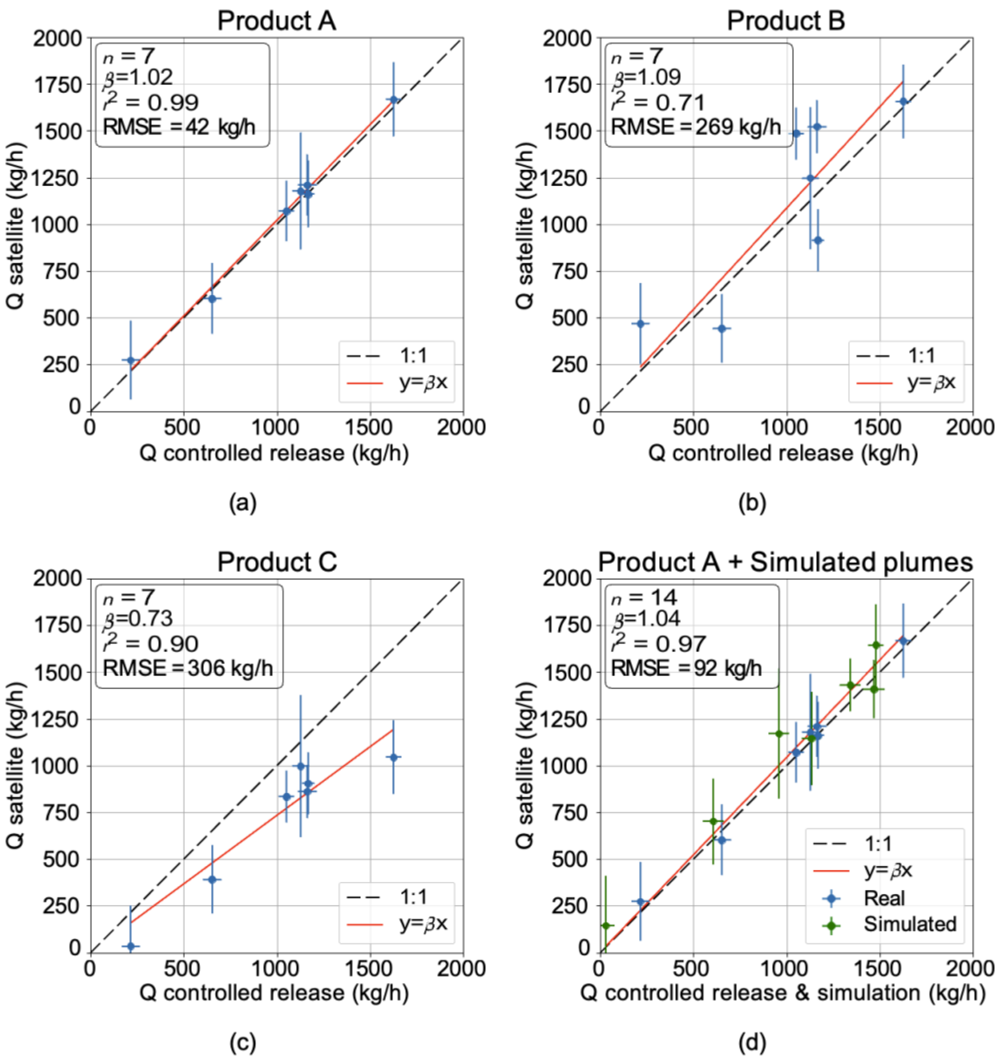


Figure 3.2: Mock-up comparison of three satellite-derived flux rate products (A, B, C) with metered flux rates (Q) from controlled methane releases. Panels (a)-(c) represent the validation of three products with different quality in terms of bias and correlation with respect the controlled release data; Panel (d) shows the validation of the product in (a), but adding further validation points based on simulated plumes (see Section 3.4).

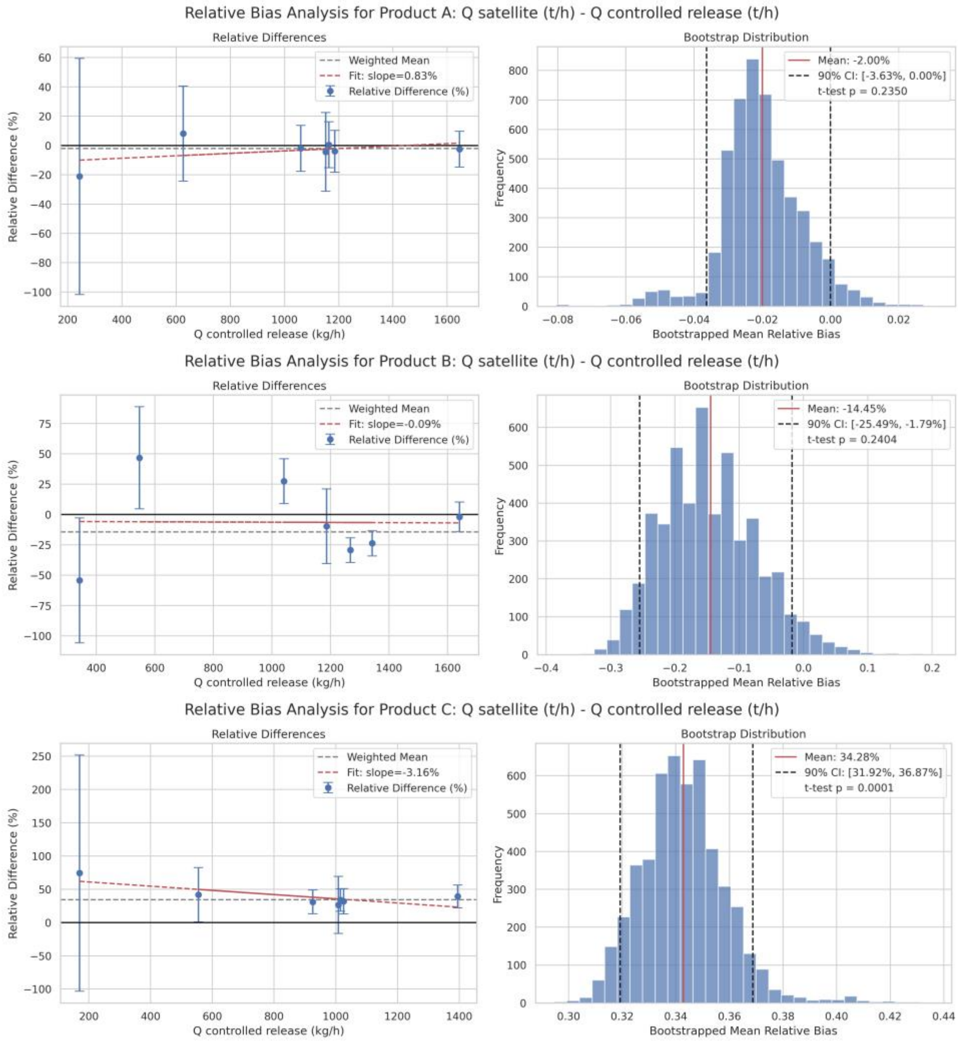


Figure 3.3: Mock-up comparison of the same three satellite-derived flux rate products (A, B, C) with metered flux rates (Q) from controlled methane releases, now using a relative bias approach. The use of the combined relative error as weight should only be considered when the errors tend to scale with magnitude (which is not the case in our examples above).

maps from which the plume has been detected, segmented and quantified. The visual inspection of those maps enables the assessment of whether retrieval noise and artifacts potentially present in the raw map would have allowed the detection of the methane enhancement caused by the controlled release in a real case scenario, in which the source location would not have been known beforehand. This step was included in the protocol of the second Stanford campaign. An example from Sherwin et al., (2024) is shown in Fig. 3.4.

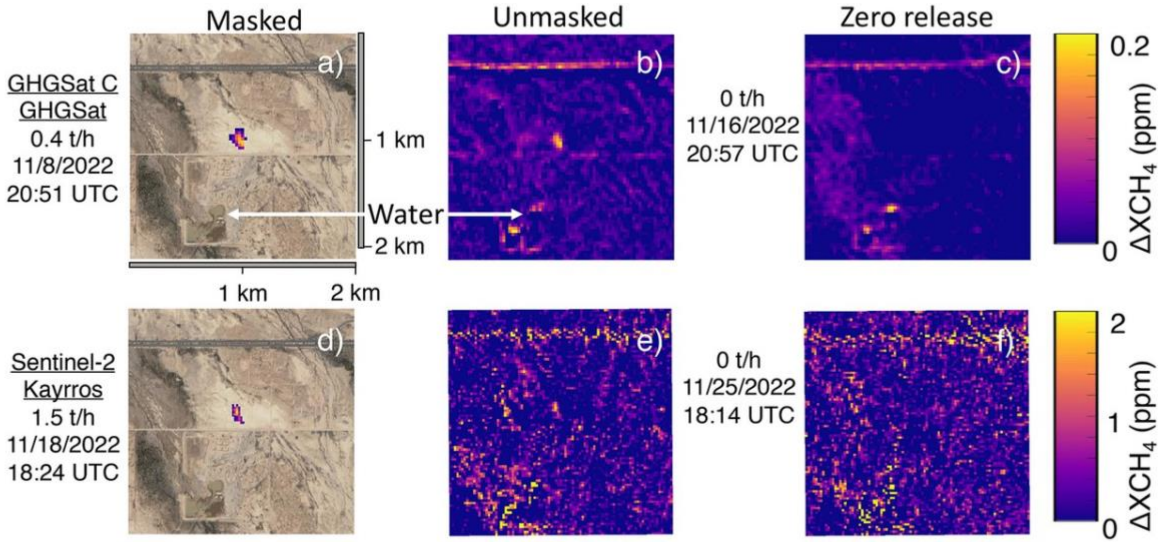


Figure 3.4: Masked and unmasked ΔXCH_4 maps showing methane plumes as well as systematic retrieval artifacts which could be mistaken by true methane enhancements during the plume detection process (reproduced from Sherwin et al., 2024).

3.3 Methodology for the intercomparison of satellite-based flux products

The quality of methane flux products is driven by both the processing methods applied for plume detection and quantification, and by the quality of the input satellite data. Accordingly, this product intercomparison task is split into inter-algorithm and inter-satellite efforts.

The intercomparison of flux products from different algorithms can be done for every satellite for which more than one processing scheme is available, whereas the product intercomparison focuses on different satellites relies on the conditional availability of coincident observations over stable sources, which are rare. This restriction can be alleviated by the use of simulated plumes to obtain a common reference basis for the intercomparison (see Section 3.4.3).

3.3.1 Intercomparison of flux products from different algorithms

The focus of this task is the relative assessment of the flux rate estimates from different algorithms. This task is closely related to the algorithm intercomparison exercise performed in MEDUSA's WP300, the main difference being that WP300 was intended to intercompare the different processing steps in the processing chain going from at-sensor radiances to flux rates. For the intercomparison of flux products from different data processing schemes (same input satellite data, but different methane concentration retrieval, plume masking, and flux rate quantification), we rely on time series of satellite acquisitions. A valuable test will be done with the time series collocated with the controlled releases, as the latter can be instrumental in the interpretation of the results from the intercomparison. Two main tests are proposed:

- **Intercomparison of satellite-derived flux rate products**: same as in the validation with controlled releases (Section 3.2), the satellite products will be inter-compared in terms of the slope of the linear regression and correlation with each other. Unlike the validation, the RMSE is not included in the intercomparison because there is no reference dataset in this case. This intercomparison will be mostly done over controlled-release sites, given that other sites of interest with different surface conditions and concentration of sources are already covered by WP300. Orthogonal regression analysis (considering uncertainties in the two datasets to be compared) will be used for the intercomparison of satellite products. Using the relative bias assessment, the individual biases are now taken as relative to the average between the two products that are compared:

$$\delta_{rel} = \frac{y-x}{m}, \text{with } m = \frac{(y+x)}{2} \quad (3.4)$$

This ensures that the obtained overall average biases are perfectly symmetrical. Also note that for the χ^2 test, the factor f with which to scale the conservative uncertainty is no longer given by y/x, but by $4|xy|/(x+y)^2$, as dictated by first-order error propagation of this symmetric relative bias. However, even greater caution is required when interpreting these values, since the compared products originate from the same satellite platform and therefore share common sources of uncertainty. In this case, the full propagated variance includes an additional covariance term,

$$\sigma^2 = \frac{16}{(x+y)^4}\left(y^2\sigma_x^2 + x^2\sigma_y^2 - 2\rho xy\sigma_x\sigma_y\right) \quad (3.5)$$

here ρ represents the correlation coefficient between the measurement errors of the two products. With sufficient data, the correlated error component may be estimated using variance decomposition techniques, or its impact assessed through a sensitivity analysis using plausible values of ρ.

- **Analysis of results supported by controlled release data**: The results from the validation exercise consisting in the comparison of satellite-derived flux products with controlled release data will be evaluated. The controlled release data will be used for the interpretation of the results from this analysis, especially in those cases in which significant deviations between different datasets are found. In particular, the controlled release data will help characterise potential biases in each dataset, which will then be incorporated in the quantitative comparison between different algorithms. We will also compare the raw methane concentrations used in the flux derivation to isolate potential difference sources.

This intercomparison process is illustrated in Fig. 3.5. The mock-up dataset consisting of 3 products from the same input satellite data which was used to illustrate the validation task (Fig. 3.2) has also been used to recreate the intercomparison process. Figs. 3.5a-b represent the linear regression analysis used for the intercomparison of products B and C with A. The scatter plots show the higher correlation between the A and C products than between A and B (r^2 of 0.9 and 0.76, respectively), although the bias between A and B (positive bias of about 7%) is smaller than that between A and C (negative bias of about 28%). A summary view of these same results is presented in the barplot in Fig. 3.5c, which would be a useful representation when more than 3 products are involved in the comparison. An alternative representation of this intercomparison exercise, but including all possible product combinations rather than selecting one as a reference, is displayed in Figs. 3.5d-e. It represents the β and r^2 of the linear regression between all three products. Again, this type of representation can be very useful when more than 3 products are included in the intercomparison.

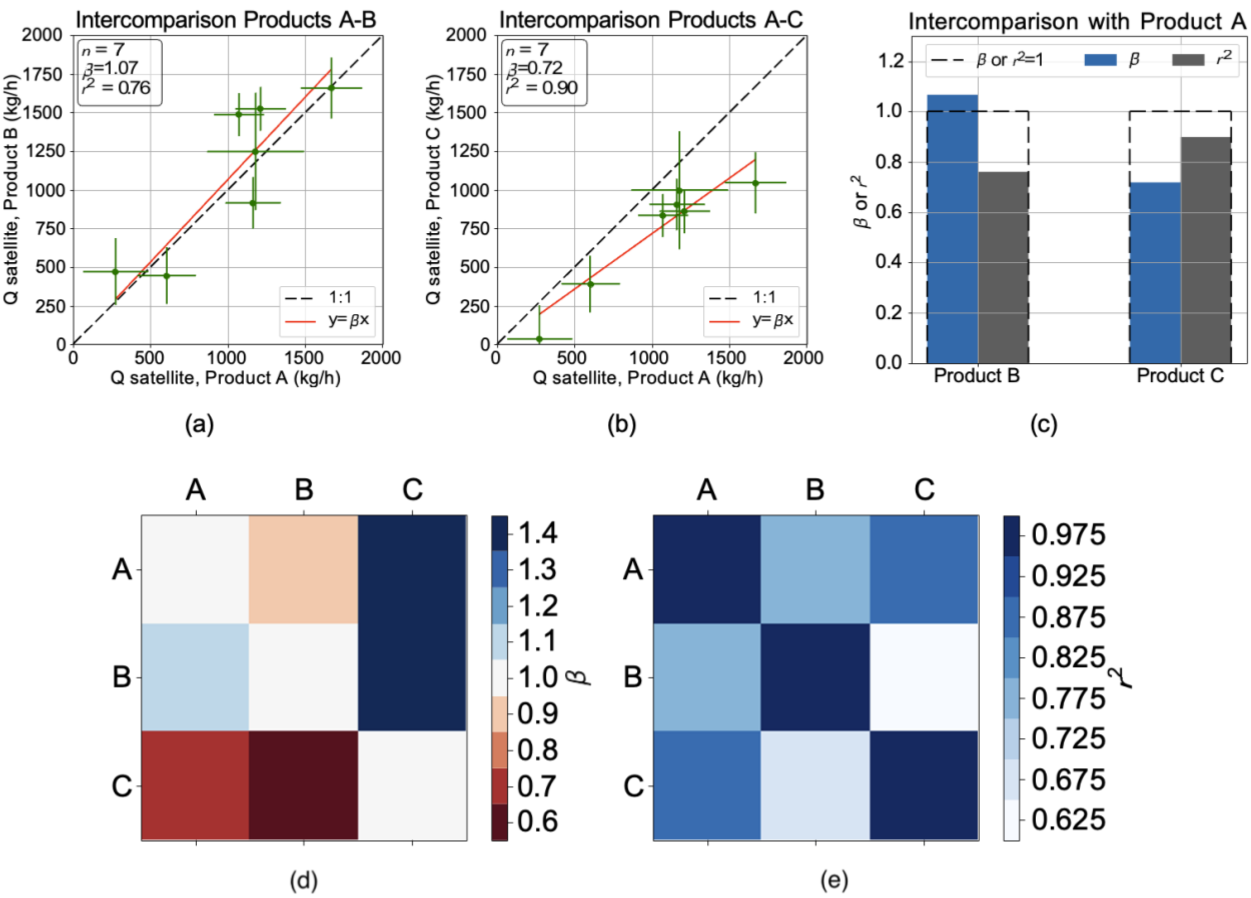


Figure 3.5: Intercomparison exercise for the same mock-up validation cases shown in Fig. 3.2. The scatter plots in (a) and (b) show the results of the intercomparison of Products B and C with A. The barplot in (c) shows a summary of the results from the comparison of A with B and C. The matrices in (d) and (e) show the β and r^2 resulting from the orthogonal regression analysis (considering uncertainties in the two datasets to be compared) of all possible intercomparisons of the three mock-up products.

The summary of the comparison between all satellite flux products and the controlled release data is displayed in Fig. 3.6. It illustrates that A would be the best performing product, as it shows the highest correlation and smallest bias with the controlled release data, both leading to the smallest RMSE. With respect to B and C, B would be the second best in terms of the RMSE, although C would be better to represent the variability in the fluxes as suggested by the higher r^2.

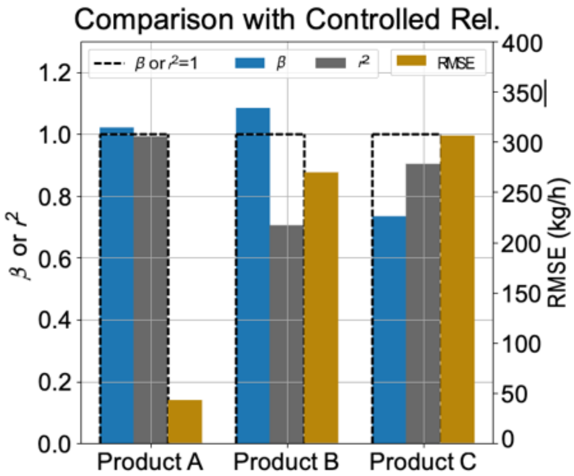


Figure 3.6: Summary of the results from the comparison of the three mock-up products with the controlled release data.

3.3.2 Intercomparison of flux products from different satellites

The intercomparison of flux products from different satellites of comparable spatial resolution requires quasi-concurrent satellite acquisitions (ideally within 2 min) to ensure that the flux rate and wind conditions at the source can be assumed to be constant for the two flux products being compared (see Appendix A for a short study of the impact of time differences on flux product intercomparisons).

Most of the hyperspectral satellites (e.g. GHGSat, EnMAP, PRISMA) need to be tasked to acquire data over a particular site, so it is difficult to obtain close-in-time acquisitions of the same source by two or more of these missions. The potential to obtain quasi-simultaneous acquisitions from high resolution missions increases if one of the high-resolution missions is Sentinel-2 or Landsat because of their continuous acquisition mode (as opposed to the “tasking mode” of the hyperspectral missions).

On the other hand, finding stable sources outside controlled release tests may be challenging due to the high emission rates which are needed to match the relatively high detection limits of space missions, especially in the case of multispectral imagers. Some landfills could play this role, but they expand over many pixels of the high-resolution instruments, and may include surface materials confounding methane retrievals, which may discourage their use as a reference and makes multispectral imagers unsuitable to detect their emissions.

We propose the following approach to overcome the challenges described previously:

- **Use GHGSat as a reference for intercomparison with the rest of the high-resolution missions:** GHGSat has the lowest revisit time and detection limit of all high-resolution missions, which makes it an ideal reference for the intercomparison with flux products from other missions. This intercomparison exercise with GHGSat as a reference will be particularly useful for the indirect validation of multispectral missions, which have relatively high plume detection limits and are therefore more difficult to validate with controlled releases. We propose to select an area of interest with a high concentration of strong active sources (e.g. around Turkmenistan's Korpezhe oil and gas field), identify the overpasses of the different missions within a given time window (e.g. two months), and task GHGSat and the other missions to acquire data with the shortest time difference possible. For this exercise, we consider that the difference in flux estimation uncertainty between different GHGSat satellites is negligible in comparison to other factors, like wind speed. Typically, GHGSat's satellite performances (detection threshold, column density precision, etc.) are always presented for the full constellation (McKeever et al., 2022).

We found out during the first validation and intercomparison cycle in MEDUSA that there are only very few match-ups between GHGSat and Sentinel-2 closer than 2 minutes in past data. Consequently, there is an on-going effort to organize co-located GHGSat observations with Sentinel-2. These must take place over persistent and strong enough sources so that (most likely) the source will be (1) emitting – as not to waste GHGSat observation budget – and (2) detectable in Sentinel-2 data. Besides, we aim to achieve temporal match-ups better than 2 minutes so that, beyond comparing emission rates, Sentinel-2 methane enhancement retrievals can also be evaluated against GHGSat detected enhancements (see Appendix A).

As the Sentinel-2 viewing geometries and orbits are fixed, the organization of such coordinated observations between GHGSat and Sentinel-2 boils down to identifying sources that can be targeted by GHGSat near the overpass time of Sentinel-2. We start from 183 sources identified in the IMEO MARS dataset that show plumes detected in band imager data. Each week, GHGSat runs orbit and target predictions for all these sources over the upcoming two weeks and, likewise, we determine the Sentinel-2 overpass times over the same time range. Same day GHGSat and Sentinel-2 overpasses with time differences lower than 2 minutes are further scrutinized. We examine the selected sources for methane emission plumes in recent Sentinel-2 data (at least past 3 months), and only sources with more than 50% persistent emissions are proposed for targeting.

Between 2025-10-01 and 2026-01-26, we identified 277 opportunities of GHGSat x Sentinel-2 match-ups better than 2 minutes, 63 were examined for recent emissions in Sentinel-2 data, and 5 opportunities were targeted, yielding 4 "clear sky" observations by GHGSat. The outcome of these targets is provided in Table 3.1.

While only few match-up opportunities occur overall, even less over persistent enough sources, this effort is valuable to continue as a tightly co-located dataset of GHGSat and Sentinel-2 observations would be a critical asset working towards an integrated satellite methane imaging intercomparison framework.

- **Identify available quasi-simultaneous acquisitions from known emission events:** In addition to the targeted acquisitions by GHGSat, co-located observations by two or more high resolution instruments may happen by coincidence. For example, Irakulis-Loitxate et al. (2022) reported methane plumes detected within a small time difference between PRISMA and Sentinel-2 on 3 July 2020 in Turkmenistan. We will explore public archives of hyper- and multispectral satellite missions in order to find pairs of datasets acquired with a short time difference. This also includes identifying additional matchups between EMIT and GHGSat and exploring the possibility of adding Tanager matchups with the Carbon Mapper team. Some examples of existing matchups between acquisitions of public high-resolution missions are shown in Table 3.1.

Table 3.1: Examples of available matchups between high-resolution missions sampling methane-relevant regions

Missions	Date (DD/MM/YY) Time (UTC)	Lat (º)	Lon (º)	Location	Comments
PRISMA-Sentinel-2	03/07/20 7:14 & 7:19	38.56	54.20	Turkmenistan	-
PRISMA-Sentinel-2	27/10/21 18:26 & 18:34	33.63	-114.49	Arizona (USA)	-
GHGSat-Sentinel-2	01/02/23 7:18:44 & 7:21:31	15.56	45.79	Yemen	-
PRISMA Sentinel-2	05/10/25 7:18:26 & 7:19:15	38.56	54.20	Turkmenistan	-
GHGSat-Sentinel-2	09/11/25 7:32:13 & 7:31:23	31.00	49.95	Iran	No GHGsat-detected plume
GHGSat-Sentinel-2	01/12/25 8:09:22 & 8:10:24	35.25	40.28	Syria	5 GHGSat plumes 0.3 – 2.4 t/hr Noisy Sentinel-2 scene (mix of desert and crops)
GHGSat-Sentinel-2	11/01/26 9:21:11 & 9:22:45	29.40	21.20	Libya	1 GHGSat plume 0.3 t/hr
GHGSat-Sentinel-2	23/01/26 7:16:41 & 7:18:43	39.47	53.79	Turkmenistan	19 GHGSat plumes 0.4 – 7.4 t/hr Thin cirrus over the scene makes Sentinel-2 retrieval difficult
GHGSat-Sentinel-2	There is an on-going effort to extend the dataset of co-located GHGSat and Sentinel-2 observations over strong methane emitting sources (see above)				
GHGSat-EMIT	NASA effort, data to be acquired				

For each of the matchups resulting from the previous approaches, we will intercompare the plume shapes and extension, and the resulting flux rates. The potential biases in each satellite data product as characterised from the comparison to controlled releases will be incorporated in the quantitative comparison between different missions.

3.4 Extension of the validation and intercomparison exercise through simulated plumes

The primary validation of methane flux rates from high-resolution missions is based on controlled release datasets, which provide real conditions, considerably lower uncertainty than the target measurement (i.e. satellite retrieval) and SI traceability to laboratory standards. Nonetheless, the range of flux rates, scenes and time periods is limited since the release campaigns require a lot of human and economic resources. Furthermore, the validation cannot trace backwards to the entire processing chain, which limits the identification of improvements/corrections. In order to complement the main validation approach, plume simulation is considered since it supports any scene, plume morphology and flux rate. In addition, it can validate the entire processing chain from radiance input to flux rate including methane enhancement map and plume mask.

Similarly to the direct validation of the satellite retrievals, we cannot directly extrapolate to other sites but make some sensitivity assessment or just a guess based on expert knowledge. However, using simulations as a validation rather than specific satellite retrievals is sensor-agnostic, cumulative across campaigns, and extensible to other gases (e.g. CO_2) offering a traceable and much cost-effective route to broad validation of satellite methane quantification.

Here we define a methodology to define such simulated products. Connecting with the validation scheme in Fig. 2.1, we detail the validation of these simulated plumes against in-situ methane releases and their use in the extension of the validation and the inter-comparison of flux products.

3.4.1 Plume simulation approach

The fundamental basis of the methodology is the use of simulated radiance products with superimposed methane plumes generated by large-eddy simulations (LES). The simulated methane plumes are converted into a spectral transmittance and embedded into real top-of-radiance datasets. This process ensures that the satellite radiance observations are not altered; thus, providing real scenes for validation.

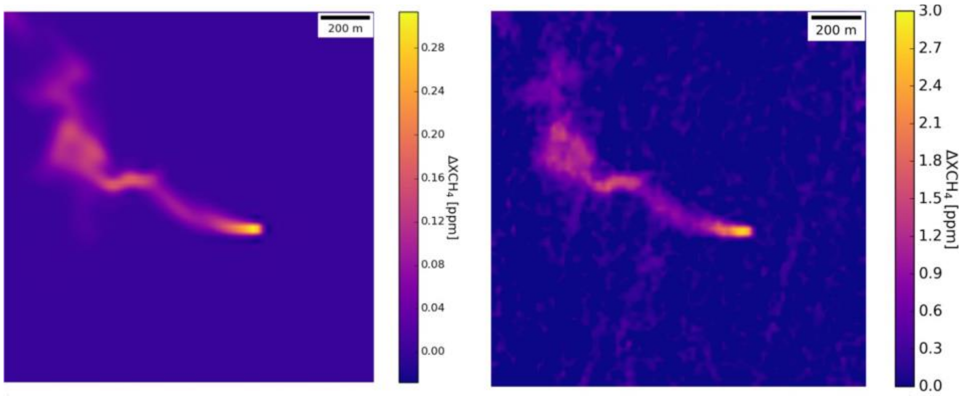


Figure 3.7: Methane enhancement map ΔX_{CH4} for (left) an input simulated plume snapshot with Q = 1000 kg/h and (right) the retrieved plume from a Sentinel-2 simulated product centered at 31.82652°, 6.15011° (Hassi Messaoud, Algeria). The original plume was scaled to Q = 10000 kg/h and resampled to 20-m resolution for the forward simulation (from Gorroño et al., 2023).

Classical data simulation approaches rely on an end-to-end framework (forward simulations and retrievals) that controls the input radiation, atmospheric parameterisation and surface characteristics as in Jongaramrungruang et al. (2021). This is a convenient method to study the sensitivity of different observation characteristics such as sensor noise, spectral resolution or aerosol optical thickness as in Yu et al. (2024). However, the more subtle specifications of the instrument (e.g., straylight or per-pixel viewing dependence) or scene characteristics might be difficult to reproduce.

The approach combining simulated plumes with independent radiance data proposed here was first used by Cusworth et al. (2019), who performed a detailed study of the detection and quantification of methane point sources for hyperspectral missions based on their spectral and noise characteristics. The spectral radiance data that they used was also simulated, though. Foote et al. (2020) added methane plumes on top of atmospherically-corrected data from the AVIRIS-NG airborne instrument. Finally, Guanter et al. (2021), Gorroño et al. (2023) and Roger et al. (2024) embedded simulated plumes on real top-of-atmosphere radiance data from the PRISMA, Sentinel-2, and EnMAP missions, respectively, which represents the same approach as proposed for this protocol.

Results from this type of simulations can be found in Fig. 3.7, which includes an example of an input simulated plume and methane enhancement map of the Sentinel-2 simulated product after flux rate scaling and resampling. In turn, Fig. 3.8 illustrates a simulated product for the MethaneAIR instrument which contains both real and simulated plumes. The examples include plumes of 500, 200 and 100 kg/h which visually exemplify their potential as realistic plumes and flexibility to model different flux rate levels.

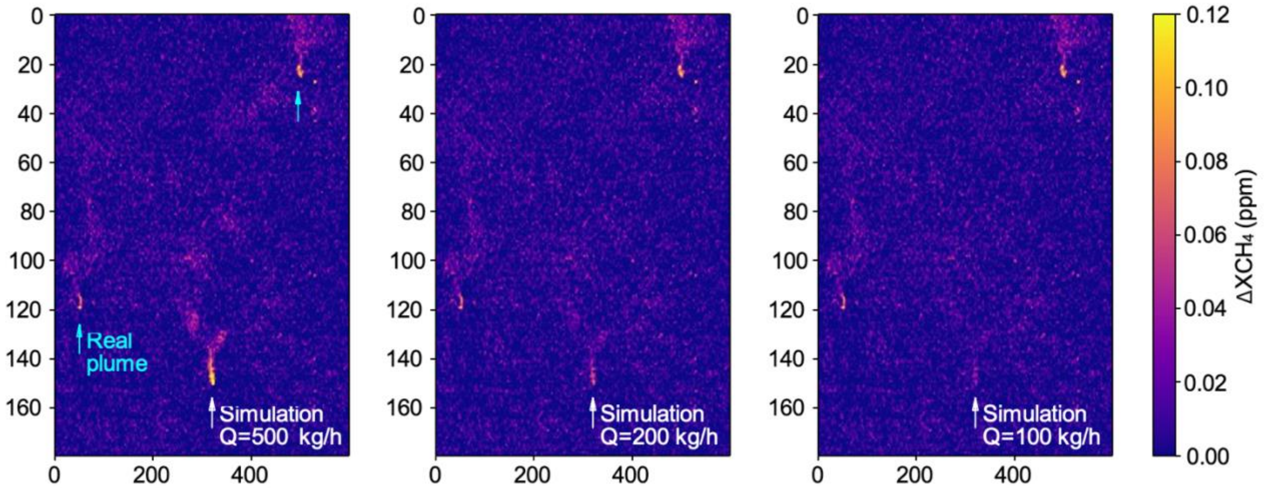


Figure 3.8: Results from end-to-end ΔXCH_4 retrieval simulations for the MethaneAIR airborne spectrometer. The simulated plume (white arrow) is scaled to 100, 200 and 500 kg/h flux rate (*Q*) values and embedded in the reference subset, which also contains real methane plumes (cyan arrows) (from Guanter et al., 2025).

A public database of transport simulations is available in Gorroño et al. 2026b They were generated with WRF-LES 4.7.0 with a spatial resolution of 20 × 20 m2, different geostrophic wind fields between 2-10 m/s (covering a wind speed range of 1-6 m/s) and a sensible heat flux H = 150 Wm−2. These simulations will be taken as a reference for the validation and intercalibration described in sections 3.4.2 and 3.4.3 respectively.

Concerning the range of emission rates for the simulations, this will be similar to what the satellites can be used for. Possibly larger because it must accommodate multiple satellites. The area of 100-1000 kg/h is already well covered, and sampling the range 1000-4000kg/h would also be important. The simulation ensemble is filtered to a U10 value close to the compared emission. We will report 2-D normalized cross-correlation between the retrieved and simulated enhancement fields to indicate the reliability of the simulation.

The comparison of plume simulations from two different transport model simulations (e.g. WRF-LES and Gaussian) for the same controlled release conditions is another test that could be performed to evaluate the robustness of the simulations, but it is considered beyond the scope of the current MEDUSA project phase.

Regarding future extensions of these simulations, there are several other parameters that could be considered. Some of them (source typology or angular parallax) are partially accomplished. However, others such as heat flux rate and vertically-resolved wind (as available from the wind lidar data acquired during the Stanford campaigns) are not yet included and are considered to be beyond the current MEDUSA nominal project phase.

3.4.2 Extended validation exercise through simulated plumes

The simulated plume products defined in Section 3.4.1 are not originally traceable to SI units.

Thus, these simulations must be connected to the in-situ releases which are, in turn, traceable to SI laboratory calibration. Thus, the simulated plumes are validated with controlled releases, so the metered flow rates become the traceable standard for both the direct validation of a satellite retrieval and the validation of simulated plumes.

Therefore, as a first step, we will define simulated products at the areas and dates from the release date campaign. Some of the scientific efforts will be based on defining improved simulations that consider the specific conditions in the boundary layer and surface of the release campaign as explained in Section 3.4.1. The simulated products cannot be placed at the exact same time and location of the methane release. However, the simulated products will be generated at 1) opposite direction of the real methane release, 2) nearby locations of the methane release and, more importantly 3) at the satellite overpasses within or close to the campaign dates with no methane emission declared.

The expected result of this alternative validation with the simulated products is similar to that presented in Section 3.2 and illustrated in Fig. 3.2. The comparison of the validation results between those of real satellite products and simulated products will provide an indication of the validity of the simulations to reproduce realistic releases. In simple mathematical terms this difference ε_{sim} can be described as:

$$\varepsilon_{sim} = \beta_{sim} - \beta_{sat} \qquad (3.6)$$

where β_{sim} and β_{sat} represent the slope of the validation curve (see Eq. 3.1 and 3.6) for the simulated products and satellite products, respectively. The comparison of the fitted curves (for both the simulations and satellite acquisitions) assumes that a large enough database of comparisons has been accumulated so that results in an uncertainty in β_{sat} that is considerably low with respect to the expected differences. Anyhow, we will consider the uncertainty over the fitted curve by propagating the individual comparison uncertainty (both reference and satellite observation). This will serve to identify whether disagreements in ε_{sim} are in agreement or not. We must also consider that this will not be specific over a single satellite but for any satellite comparison leading to a consolidation of any potential biases due to minimisation of uncorrelated errors.

Once these simulations are anchored to in-situ methane releases, they will be used to extend the original validation to other scenes, flux rates or plume morphology. The plume simulations will be those described in Section 3.4.1 and, if needed, these simulations will be adapted to the specific site and observation conditions. The comparison methodology is the same as described for the operational validation in Section 3.2, but with the simulated plumes taken as the reference measurement.

Because the use of simulations implies larger number of points and the possibility to trace back the different levels of processing, this will lead to a richer validation analysis. For example, it will be possible to study the validation depending on the flux rate level and the probability of detection. The latter will be part of a study in WP400 restricted to EnMAP and Sentinel-2.

Finally, it will be possible to extend validation to other sites such as those where we expect no controlled releases or are technically complex to implement. This implies, however, that the simulations perform similarly at the controlled release site where they have been validated and at the site to which the validation is to be extended. Although we do not see a priori any reason why this would not be the case, it could still add some uncertainty for this particular validation exercise. We will investigate potential biases in the simulations associated to site-dependent factors such as surface albedo and winds.

3.4.3 Intercomparison of flux products from different satellites anchored to simulated plumes

For the identified cases where two missions overpass an active methane source within a small time difference, we will compare the corresponding flux rate estimates (Section 3.3.2). While near-coincident plumes will be very useful to anchor the inter-comparisons, they will be limited in terms of number and analysis, and it will be very difficult to disentangle the potential sources of discrepancy.

Therefore, simulated products as defined in Section 3.4.1 will be used for additional inter-comparisons. These simulations enable inter-comparison between non-simultaneous acquisitions by different missions, which can result in a large number of potential comparisons and is generally better than assuming stable emissions for acquisitions close in time. It is expected that most error sources arising from the simulation itself are cancelled out due to the relative nature of the comparison. In addition, by using the exact same plumes for the different satellites, it will be possible to inter-compare not only the flux products but the entire processing chain (e.g. methane enhancement map) similarly to the validation approach in Section 3.4.2.

4 Validation and intercomparison of methane flux products from global flux mappers

4.1 Review of resources for validation of flux products from global flux mappers

Global flux mappers (FM) cannot be validated using controlled release data due to their high detection limits and large footprints. Consequently, “true” direct validation is not possible, and alternative methods must be developed and explored (see Fig. 4.1). This Section also covers the indirect validation of the Sentinel-3/SLSTR and GOES/ABI missions. Although these missions are not global flux mappers, their coarse resolution prevents them from being validated through controlled releases. Instead, their indirect validation will follow a methodology akin to what is employed for global flux mappers. The first method used to gain confidence in the flux products (detection and emission quantification) is intercomparison of products derived from flux mappers by different data providers. Since Kayrros is the sole general party for Sentinel-3, no such intercomparison will be carried out for this mission. In the first 12-month cycle the only flux mapper data is from the Sentinel-5P TROPOMI instrument. We look at the same observations but either using a different L2 XCH4 column product and/or different detection and emission quantification algorithms. This work connects to the comparisons done within WP300 for the first cycle. For the second 12-month cycle in the project, we will include more recent Sentinel-5P TROPOMI data as well as synthetic observations to help better explain product-to-product differences.

Indirect validation and intercomparison of flux products from global flux mappers

Intercomparison of flux products from different algorithms

- To evaluate the relative performance of flux products obtained from the same satellite by different data processing schemes

Comparison with ‘validated” flux products from high resolution missions

- Flux products from high-resolution systems validated with controlled releases are used as the validation reference for coarse resolution flux products

Figure 4.1: Summary of tasks for the validation and intercomparison of flux products from global flux mapping satellites.

The second more experimental method involves comparing FM data with emission products from high-resolution satellites that have been validated - to some extent - using controlled releases. This approach is still to be developed within the MEDUSA project and presents two main challenges. First, it is essential to select isolated sites to ensure no other significant sources are contributing the FM-detected plumes. Second, temporal mismatches between flux mapper acquisitions and high-resolution data must be accounted for, or sources must be very stable, as the time-scales of high-resolution plumes are shorter (~1h) than FM-detected plumes (~6h).

Candidates for these comparisons are large and isolated events, if possible with extended durations. Guanter et al., 2024 reported on a well blowout in Kazakhstan's Karaturun East oil field using TROPOMI and various high-resolution missions, including some that have been previously evaluated using the Stanford controlled releases (PRISMA, EnMAP, Sentinel-2, Landsat 8/9). However, blowout events have temporally changing emissions, which requires either almost simultaneous observations or sufficient observations to be able to account for those temporal changes. For example, Cusworth et al., 2021 combined observations from several satellites, including TROPOMI and PRISMA, to characterize emissions from a 2-day leak event due to a gas well blowout in the Eagle Ford Shale (USA). However, the space-based sampling for this event is scarce relative to the change in emissions and as such the emissions from various satellite instruments cannot be used for comparison evaluations. Other candidates for comparisons could be well-isolated (as stable as possible) persistent sources, possibly venting shafts from underground coal mines assuming all emissions are captured with the high-resolution satellites as flux mappers integrate all emissions in their footprint. Currently, Sentinel-3 and GOES are primarily employed in a tip and cue approach with TROPOMI (Pandey et al. 2023 and Watine-Guiu et al., 2023). Due to its better spectral resolution, TROPOMI will serve as reference data in addition to high-resolution missions, enabling more relevant match-ups to be obtained.

In this second validation cycle we will expand upon the use of medium-resolution (~500m) band imagers. These imagers (Sentinel-3, VIIRS) provide an opportunity to form a bridge between high-resolution and flux mapper observations. VIIRS was not included in the first validation cycle but is added now as it is especially promising in this regard since the instrument on Suomi-NPP observes the atmosphere within a few minutes of the TROPOMI overpass. De Jong et al. (2025) first showed that VIIRS can be used to detect large, concentrated methane plumes and demonstrated the direct overlap between methane plume concentrations observed by TROPOMI and VIIRS (Figure 4.2). Once a large methane plume is detected by TROPOMI it can be followed up with VIIRS if the emitted mass is large enough and the surface albedo allows detection using the limited spectral resolution of the band imager. This approach mainly holds value for large transient emission events where other (targeted) observations are often not available. For these events, the multiple daily overpasses by VIIRS and Sentinel-3 also allow us to track methane plumes over time (Figure 4.2). Sentinel-3 observes close in time to Sentinel-2, hence a plume detection by TROPOMI can be linked to Sentinel-2 using VIIRS and Sentinel-3 as a bridge. Matchups between TROPOMI, medium-resolution band imagers, and high-resolution band imagers are included in Table 4.1. Additional matchups can be identified and added during the validation cycle. SRON will also process the data from instruments it does not cover for other sections of the project for this evaluation. While the focus of MEDUSA is on emission rates, we will also perform a comparison of detected methane enhancement as it enables us to better understand observed differences and identify possible retrieval biases.

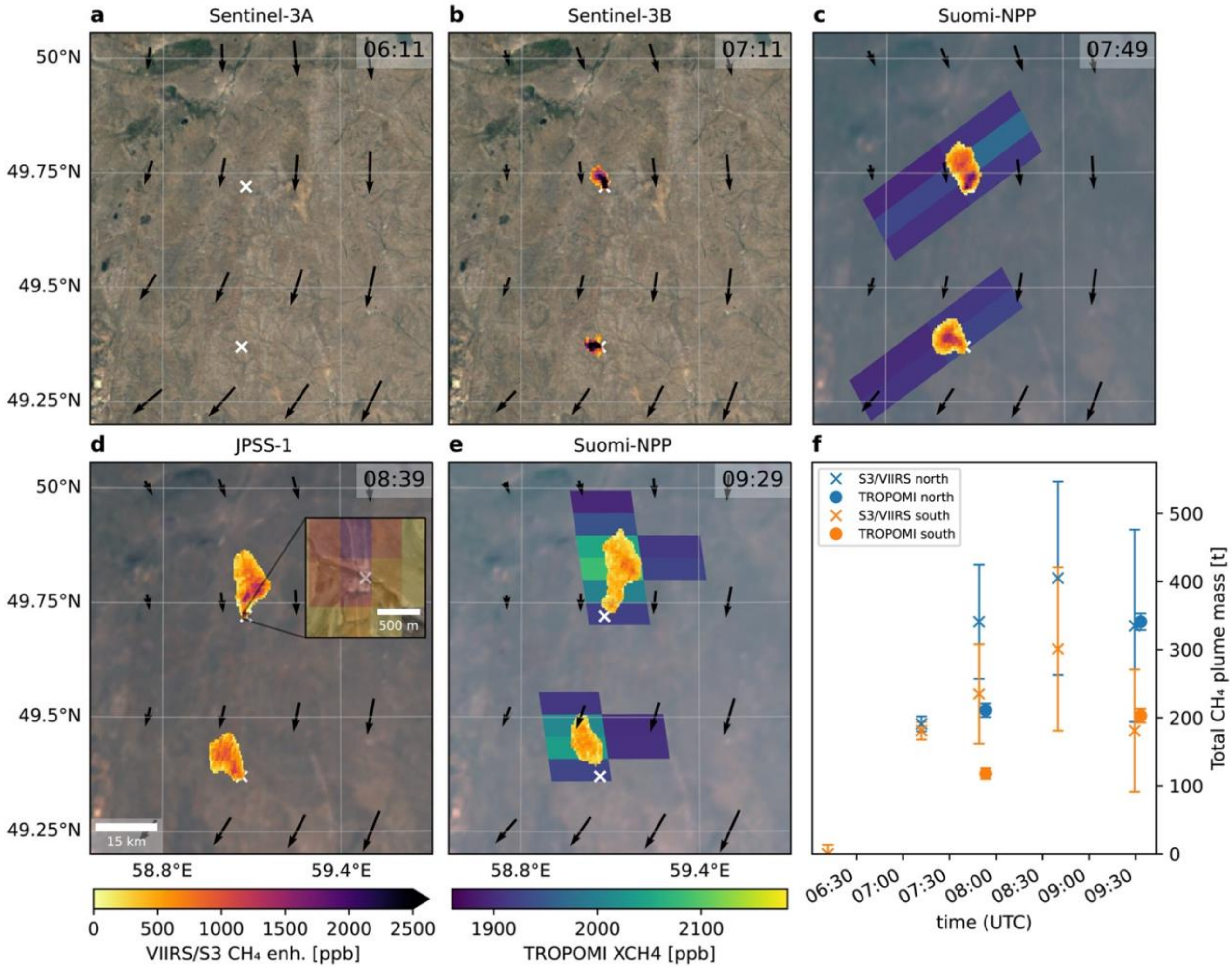


Figure 4.2: Taken from de Jong et al. (2025), showing the overlap between Suomi-NPP VIIRS and TROPOMI methane plume observations. The event was also detected using Sentinel -2 (Schuit et al., 2023). Caption extracted from de Jong et al. (2025): Gas release at two gas pipeline block valve stations in Kazakhstan on 14 May 2021. (a) Sentinel-3A overpass before emissions started. (b) Sentinel-3B overpass after the start of emissions (background (a) and (b) Google Maps imagery (Google, 2024)). (c–e) JPSS overpasses. For the Suomi-NPP overpasses, the corresponding TROPOMI pixels of the plume are also shown. Background images are VIIRS level-1-derived RGB images of the respective overpasses. (inset d) Zoom of the northern source of the northern plume overlayed over a Google maps background (Google, 2024). (f) Timeline of the total plume mass estimates. Arrows in panels (a–e) correspond to the 100 m ERA5 wind, interpolated to the time of overpass, with arrow length corresponding to the transport during 1 hour.

An additional approach to the validation can come in the form of 'coordinated' releases with well-known emission rates. These vary from controlled releases as they are part of "regular" oil and gas operations and do not have metered flow rates but are instead characterized based on available (engineering) information provided by the emitter. An example of this kind of release is the 'Very Large Methane Release' (VLMR) that was coordinated by NOAA with a US gas transmission company in 2024 (He et al., 2026). A gas pipeline blowdown emitting around 350 tons of methane from two release points over the course of an hour was coordinated with ground-based, aircraft, and targeting satellite observations. In addition, instruments with regular coverage such as TROPOMI, VIIRS, and GOES observed the event. The large emission rate of this type of event provides a validation data point on a very different emission magnitude than typically covered by controlled releases and provides opportunity to evaluate both total detected mass as well as emission rates. The VLMR event (He et al., 2026) will be included in the second validation cycle and similar events can be added as they become available.

A critical difficulty of comparing FM-derived emissions to high-resolution emission estimates is the large gap in spatial resolution, leading to large time scale differences in the observed plumes. On the top of this, real-life data can add more complications e.g. different sources emitting at the same time. Consequently, a final exploratory way of linking FM-detected emissions with high-resolution data could leverage atmospheric transport simulations. These could be performed using WRF-LES embedded in WRF simulations, so that the emissions of a given source can be modelled consistently across spatial scales. Such modelling efforts could help tie emission estimates across spatial scales in a simplified synthetic environment. The feasibility of this work will be assessed during the validation cycle.

4.2 Intercomparison of flux products from different algorithms

As in Section 3.3, the overall objective of the product intercomparison task is to evaluate the relative performance of different flux mapper products, considering both their retrieval and quantification methods. The intercomparison will be divided into two parts, each utilizing their own datasets:

- **Intercomparison of FM flux rate products, based on a common input dataset defined within WP300**. This relates to both the global detections as well as the target regions. In WP300 this was already done for the year 2021 which results were used in the 1st cycle. For the second cycle, we will use a more recent year for these intercomparisons. The products will be compared in terms of the slope of the linear regression and correlation with each other. RMSE will not be included as there is no reference data.
- **Intercomparison of FM flux rate products or of FM plume mass at the same overpass time (e.g. relevant for GOES)**, at location/times covered by the high-resolution satellite intercomparison, using the defined dataset in Section 4.3 of the present document. The metrics here are the slope, the coefficient of determination (r^2), and the RMSE.

This intercomparison process will be the same as the one illustrated in Fig. 3.5 which shows the process for point-source imagers.

4.3 Comparison with “validated” flux products from high resolution missions

The aim is to compare products from global flux mappers with those from high-resolution missions that have been validated using controlled releases.

The first step involves constructing a relevant match-ups database. Spatially, the satellite instruments need to be observing the same isolated point-source, where co-located diffuse sources should be avoided. Temporally, we need to be as close as possible, as emissions will vary in time. We also need to obtain longer time series where possible, to have an indication of variability. As an experimental approach, the use of observations with higher revisit frequencies (e.g. GOES) may also help as it can provide emission estimates at different timestamps, and as such be used to connect satellites with different overpass times.

Table 4.1a: TROPOMI – high-resolution match-up dataset built from UNEP's IMEO MARS database

See Appendix file: *phase2_tropomi_imeo_mars_match-ups_extVIIRS_rev2_commented.xlsx*

Table 4.1b: TROPOMI – GHGSat match-up dataset

See Appendix file: *phase2_tropomi_ghgsat_match-ups_commented.xlsx*

Table 4.2: Match-ups that will be explored to intercompare Sentinel-3 SLSTR emissions estimates.

Area	**Latitude (deg)**	**Longitude (deg)**	**Source type**	**Surface type**	**Persistence**	**Date (S-5P orbit)**	**Observed so far by**	**References**
Algeria	31.92	6.17	Oil/gas	desert	event	2020/03/01	TROPOMI, Sentinel-2, Sentinel-3 SLSTR	Pandey et al. 2023
USA	32.13	-103.71	Oil/gas	desert	event	2024/02/19	TROPOMI, Sentinel-2, Sentinel-3 SLSTR	
Algeria	28.68	7.19	Oil/gas	desert	event	2020/10/14	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	40.01	60.94	Oil/gas	desert	event	2020/02/28	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	36.48	61.62	Oil/gas	desert	event	2021/04/10	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	38.58	62.73	Oil/gas	desert	event	2022/01/07	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	39.98	58.64	Oil/gas	desert	event	2023/03/24	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	38.57	62.79	Oil/gas	desert	event	2023/01/10	TROPOMI, Sentinel-3 SLSTR	
Kazakhstan	44.86	54.35	Oil/gas	desert	event	2020/09/23	TROPOMI, Sentinel-3 SLSTR	
Kazakhstan	43.51	75.55	Oil/gas	desert, montaneous	event	2021/07/24	TROPOMI, Sentinel-3 SLSTR	
Libya	28.7662	20.9111	Oil/gas	desert	event	2021/02/25	TROPOMI, Sentinel-3 SLSTR	
Russia	58.6119	40.1056	Oil/gas	vegetation	event	2024/05/09	TROPOMI, Sentinel-3 SLSTR	
Kazakhstan	41.7098	67.099			event	2025/03/18 (38490)	TROPOMI, Sentinel-3 SLSTR	

Turkmenistan	38.5396	53.9401			event	2025/07/28 (40363)	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	38.5396	53.9401			event	2025/07/28 (40363)	TROPOMI, Sentinel-3 SLSTR	
Turkmenistan	38.5675	62.6976			event	2025/10/01 (41285)	TROPOMI, Sentinel-3 SLSTR	
Syria	35.3756	40.4984			event	2025/10/14 (41470)	TROPOMI, Sentinel-3 SLSTR	

Table 4.3: Match-ups that will be explored to intercompare GOES-16 emissions estimates indirectly.

Area	Latitude (deg)	Longitude (deg)	Source type	Surface type	Persistence	Date	Observed so far by	References
Mexico	26.0941	-104.3308	Oil/gas	desert	event	2019/05/12	TROPOMI, GOES Sentinel-2, Sentinel-3 SLSTR	Watine-Guiu et al. 2023
Idaho	43.0478	-115.6902	Oil/gas	desert	event	2023/10/12	TROPOMI, GOES	Groshenry et al., 2024
Permian	31.2715	-103.0984	Oil/gas	desert	event	2023/10/31	TROPOMI, Sentinel-3 SLSTR, GOES	
Permian	31.8542	-101.8048	Oil/gas	desert	event	2023/11/22	TROPOMI, Sentinel-3 SLSTR, GOES	
Permian	32.2407	-101.9394	Oil/gas	desert	event	2024/01/15	TROPOMI, Sentinel-3 SLSTR, GOES	
Indiana	41.0037	-87.5747	Oil/gas	vegetation	event	2020/06/14	TROPOMI, GOES	
Indiana	40.5973	-88.1488	Oil/gas	vegetation	event	2020/06/14	TROPOMI, GOES	
Argentina	-36.1390	-67.0169	Oil/gas	desert	event	2020/05/30	TROPOMI, GOES	
Arkansas	33.7358	-93.8809	Oil/gas	vegetation	event	2023/12/26	TROPOMI, GOES	

In Table 4.1a, we provide a revised version of the matchup database between TROPOMI and high-resolution emission estimates (excluding GHGSat). This new version has been compiled using the SRON TROPOMI plume product spanning 2023-01-01 – 2026-02-14 and the 2026-02-18 version of the IMEO MARS dataset, spanning 2021-05-05 – 2026-01-12. To compile this update, we explored all match-ups within 6 hours and 50 km, up from 10 km for Phase 1 to allow for uncertainty in the TROPOMI plume source estimation. The revised Table 4.1 now proposes a tiered match-up dataset that includes:

1. a core ‘isolated match-up’ subset (N=12). It contains all the match-ups that show no other high-resolution source from the IMEO MARS dataset within 50 km from the match-up itself, excluding city areas (Charlotte, Casablanca and Karachi), and only considering the instruments that can be validated against controlled releases (Sentinel-2, Landsat, PRISMA, EnMAP and EMIT). This isolation criteria is very strict and for example does not allow match-ups from the Kazakhstan blowout (Guanter et al., 2024) to be included. Because it is geographically unambiguous (up to the current completeness of the IMEO MARS dataset) and determined with objective filters, this core ‘isolated match-up’ subset is currently the best one to tie TROPOMI plume quantifications to high-resolution emission estimates validated with controlled releases, and thus build the traceability framework from controlled releases to TROPOMI.
2. an extended ‘comparison’ subset (N=37). It has been manually curated, examining all the TROPOMI to high-resolution match-ups. This non-blind selection considered both the density of other sources around candidate match-ups and the overlap between SRON’s TROPOMI and IMEO’s high-resolution quantification ranges. It includes sources that are more or less isolated but for which SRON and IMEO quantifications somewhat match. While not objectively selected as to build the MEDUSA traceability framework, exploring these cases will bring insights on the conditions related to high-resolution detected emission quantifications somewhat matching TROPOMI plume quantifications. This extended comparison subset also includes VIIRS and Sentinel-3 IMEO plumes that are related to emission events also captured by one of the controlled release-validated instruments (Sentinel-2, Landsat, PRISMA, EnMAP and EMIT). It is also important to note that UNEP IMEO’s quantifications are not part of the MEDUSA project, and that only MEDUSA partners’ quantifications will be considered in the intercomparison exercise.

In Table 4.1b, we provide the complementary match-up database between TROPOMI and GHGSat. It has been compiled using the SRON TROPOMI plume product spanning 2023-01-01 – 2024-12-31 filtered for oil, gas and coal-related emissions, and the GHGSat catalog dataset. To compile this dataset, we explored all match-ups within 6 hours and 50 km to allow for uncertainty in the TROPOMI plume source estimation. These filters yielded 43 TROPOMI plumes matching 47 GHGSat plumes. The final selection was manually curated, picking match-ups as geographically isolated as possible (given the dataset of IMEO-detected sources). It did not consider emission rates as GHGSat did not share those before data selection.

In the same manner, Tables 4.2 and 4.3 provide a first list of events that will be investigated for the comparison between Sentinel-3 or GOES with high-resolution missions or TROPOMI. While the most logical comparison for Sentinel-3 would be with Sentinel-2 emission estimates, only two such match-ups have been identified so far. To expand the dataset with more locations, surface types, dates, and emission rates, we included TROPOMI as a reference. Given the low methane sensitivity of Sentinel-3 and GOES, this analysis is limited to oil and gas events. The short time differences between TROPOMI and GOES (a maximum of 5 minutes) and between TROPOMI and Sentinel-3 (less than 2 hours) make the assumption of stable emissions more realistic for oil and gas sources. In the case of GOES, the comparison is more complex, and two approaches will be evaluated: either comparing the TROPOMI plume mass with the GOES plume mass at the closest timestamp, or using the GOES emission rate based on all timestamps compared to TROPOMI emission rate. The comparative analysis will utilize the match-ups database and will be based on linear regression fits of estimated emission rates and/or detected-plume mass if relevant (e.g. for GOES). The fits will employ orthogonal distance regression to account for uncertainty in both datasets. In this analysis, flux rates from high-resolution missions will be plotted on the x-axis, while flux rates from the flux mapper to be compared will be plotted on the y-axis. If sufficient match-ups can be found, three metrics will be computed for these intercomparison: the slope, the coefficient of determination (r^2), and the RMSE (taking the high-resolution estimates as the truth), consistent with the metrics used for the validation of high-resolution instruments. In addition, detailed analyses will be needed when we see differences.

As mentioned in section 4.1, the 2nd cycle will be expanded upon the use of medium-resolution (~500m) band imagers. These imagers (Sentinel-3, VIIRS) provide an opportunity to form a bridge between high-resolution and flux mapper (TROPOMI) observations. VIIRS is especially promising in this regard since the instrument on Suomi-NPP observes the atmosphere within a few minutes of the TROPOMI overpass. This approach mainly holds value for large transient emission events (large enough for VIIRS to be detected and) where other (targeted) observations are often not available. For these events, the multiple daily overpasses by VIIRS and Sentinel-3 also allow us to track methane plumes over time. Sentinel-3 observes close in time to Sentinel-2, hence a plume detection by TROPOMI can be linked to Sentinel-2 using VIIRS and Sentinel-3 as a bridge. Matchups between TROPOMI, medium-resolution band imagers, and high-resolution band imagers are included in Table 4.1. Additional matchups will be identified and added during the validation cycle (see Section 5.2.3). We will also perform comparisons of detected methane enhancement as it enables us to better understand observed differences and identify possible retrieval biases.

Additionally, for each satellite dataset used to calculate flux rate values for intercomparison, it is required to provide the raw methane concentration maps. These maps should indicate where the plume has been detected, segmented, and quantified (see Fig. 4.3).

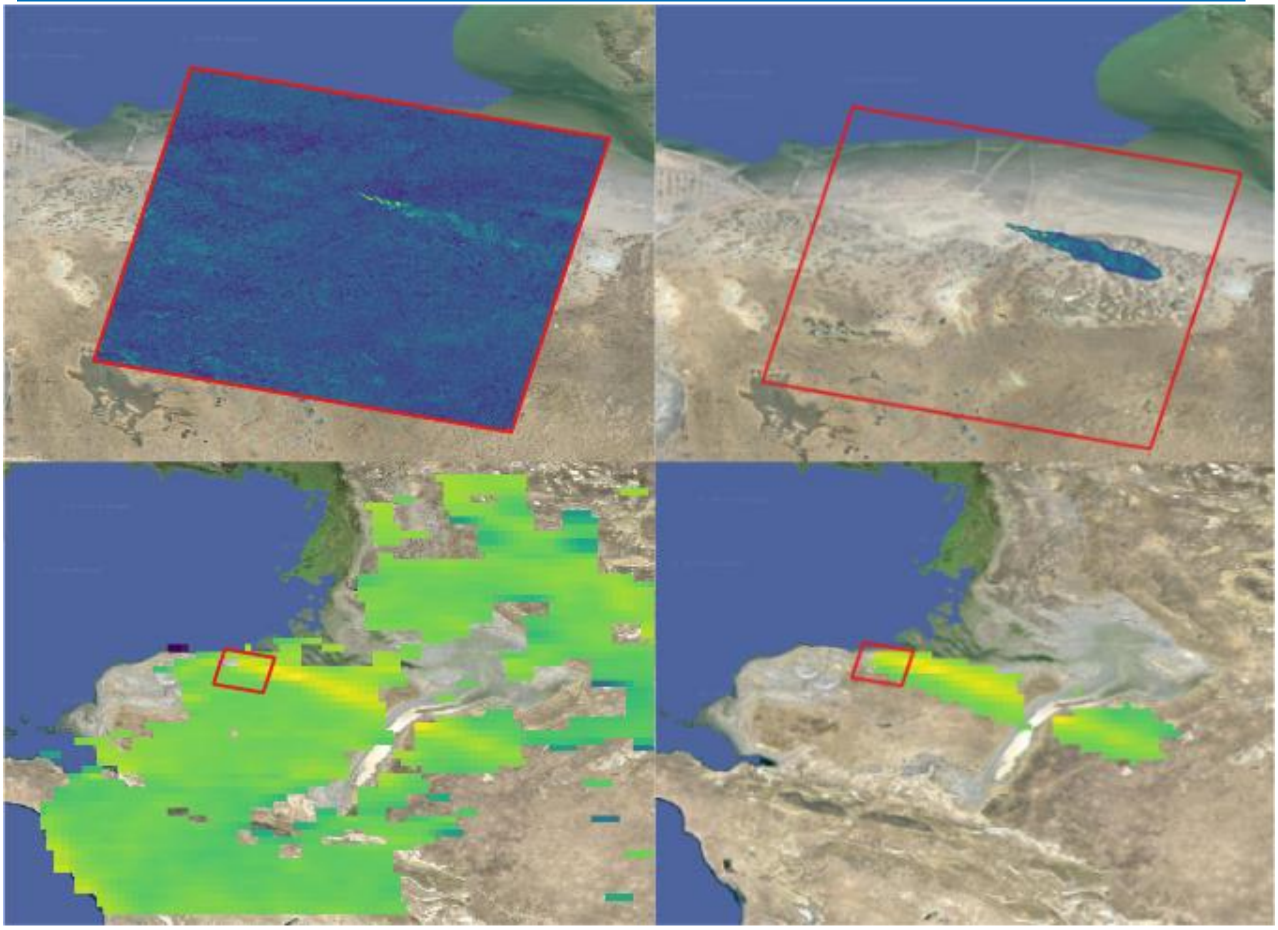


Figure 4.3: Masked and unmasked concentration maps for the match-up included in Table 4.1. The first line of images corresponds to PRISMA and the second one to TROPOMI. The left images show the full data field, the images on the right only show the masked plume. The red box marks the full PRISMA observation and is included in all panels.

5 Tasks for MEDUSA's second cycle of validation and intercomparison methane flux products

Details on the implementation of the different tasks for MEDUSA's second validation and intercomparison cycle are provided in the remaining of this Section.

These tasks will be performed during MEDUSA's WP220 *System Implementation and Demonstration*. The main results will be included in the second version of the PVIR. The second Phase of MEDUSA's WP220 will see increased interactions between data providers and the partner implementing the intercomparison framework, with regular (e.g. bi-weekly) meetings to support intercomparison results with expert knowledge and – if necessary – complementary tests to bring more understanding into the origins of possible differences.

5.1 Tasks for MEDUSA's second validation and intercomparison cycle of high-resolution methane flux products

Based on the discussion in Section 3, the proposed tasks for assessing the quality of flux rate products from high-resolution (HR) missions are as follows:

- **Task HR-1**: Validation of flux rate estimates through comparison with controlled releases
- **Task HR-2**: Intercomparison of flux products from different algorithms
- **Task HR-3**: Intercomparison of flux products from different satellites
- **Task HR-Scientific**: Demonstration of the extended validation and intercomparison approach using plume simulations

5.1.1 Task HR1-: Validation of flux rate estimates through comparison with controlled releases

For the second validation and intercomparison cycle in MEDUSA, we will use the existing datasets from high resolution satellites and controlled releases generated during the available Stanford campaigns (2021, 2022 and 2024/2025). We will re-evaluate the data from the 2021 and 2022 campaigns, even though they were already used in MEDUSA's first validation and intercomparison cycle. The number of available acquisitions from hyperspectral missions during the 2024-2025 experiment are shown in Table 5.1. It shows that the amount of data from the 2024-2025 campaign is substantially higher than what was available from the two previous campaigns together. This higher number of controlled release match-ups will improve the robustness of the statistical analysis with respect to that from MEDUSA's first cycle.

Table 5.1: Available match-ups of satellite-based imaging spectrometers and controlled releases from the 2024-2025 Stanford campaign, compared with those of the 2021 and 2022 campaigns. These numbers include releases that occurred during high cloud cover or that might be excluded due to internal quality check (Reuland et al., 2026). The controlled releases for Sentinel-2 and

Landsat during the 2024-2025 campaign are not considered for analysis because the emission rates below 1200 kg/h in this campaign were deemed too low for reliable quantification with those missions.

	GHGSat	EnMAP	PRISMA	EMIT	Sentinel-2	Landsat
2021 & 2022	15	2	11	0	13	10
2024-2025	**84**	**41**	**26**	**7**	**71**	**35**

The EnMAP, PRISMA and EMIT data will be processed by the MEDUSA teams involved in hyperspectral data analysis. In order to keep the analysis simple, only one algorithm will be used by each team for the processing of data from each satellite. However, it must be noted in Table 5.1 that the 50-1200 kg/h range of emissions in this experiment (see Section 3.1) ranges from well below to close to the Sentinel-2 and Landsat multispectral system detection thresholds.

The Sentinel-2 and Landsat data acquired during the 2024-2025 campaign will not be included in the analysis due to too low controlled release rates compared to their detection thresholds. In particular, the peak emission rates in the 2024-2025 Stanford campaign were about 1200 t/h, with most of the releases being below 800 kg/h. These values are well below the minimum detection limits accepted for Sentinel-2 and Landsat. For example, Gorroño et al., (2023) estimated the Sentinel-2 minimum detection limit to be in the range 1000-2000 kg/h for ideal observation conditions. Some plumes below 1000 kg/h could still be detected in controlled release campaigns, where the location of the source is known beforehand. However, we consider that those plumes would not be detectable under real operation conditions, so releases of this magnitude can not serve as a reliable validation reference for Sentinel-2 and Landsat. For this reason, we will reprocess the 2021-2022 Stanford controlled releases for Sentinel-2 and Landsat under the new analysis set-up proposed for MEDUSA's 2nd validation and intercomparison cycle. We will further elaborate on this in PVIR V2, including by considering the Kayrros and IMEO analyses/data from Reuland et al., 2026; Also aiming to provide quantification of a lower bound on emission rates/methane enhancements necessary for Controlled Releases to validate S2 and L8.

The methodology described in Section 3.2 will be followed for the analysis of the flux products. The following results shall be generated:

- **Task HR-1, result #1**: scatter plots and metrics (β, r^2, RMSE, relative bias and probability values of the bias and standard deviation) from the comparison of each flux product with collocated controlled release data (similar to Fig. 3.2). Taylor diagrams could also be used for the visualization of the results.
- **Task HR-1, result #2**: raw XCH4 or ΔXCH4 maps for the satellite acquisition containing the smallest flux-rate plume of those used in the comparison with controlled release data (similar to Fig. 3.4).

These quantitative outputs will be combined with an analysis of the main results and the interpretation of the identified differences between flux products when possible.

The comparisons between in situ and satellite-based flux rates may consider both reanalysis and in situ wind measurements to better understand the impact of wind data in the flux rate estimates.

Data products for all satellites will be compared individually against controlled releases. To prepare for the comparison against flux mappers, data providers will examine individual validation results, and data products with similar results will be aggregated together (e.g. averaged) in order to simplify the intercomparison between different satellites. Satellite-wise results will then be re-examined in order to discuss high-resolution satellite grouping/selection to carry the controlled release reference to Flux mappers/TROPOMI

5.1.2 Task HR-2: Intercomparison of flux products from different algorithms (high resolution)

The data from high resolution systems acquired over controlled releases from the 2025 Stanford campaign will be used in this task. Following the methods described in Section 3.3.1, the following results shall be obtained:

- **Task HR-2, result #1**: matrix plot showing the r2, β, relative bias and probability values of the bias and standard deviation from intercomparing all combinations of flux products obtained by different algorithms for the same satellite (similar to Fig. 3.5d-e).
- **Task HR-2, result #2**: barplot including r2, β, and RMSE from the intercomparison of controlled releases with all the flux products obtained by different algorithms from the same satellite (similar to Fig. 3.6).

The same exercise could be performed over other study sites, although we reckon that part of this algorithm intercomparison work was already partly carried out in MEDUSA’s WP300.

No algorithm intercomparison will be possible for GHGSat, since only the GHGSat team will process the data.

5.1.3 Task HR-3. Intercomparison of flux products from different satellites

As discussed in Section 3.3.2, the intercomparison of flux products from different satellites is conditioned by the availability of data. For this second cycle in MEDUSA, we anticipate to use the available match-ups between high-resolution satellites listed in Table 3.1 as a baseline.

In parallel, we are gathering a database of match-ups between GHGSat and high resolution satellites for analysis during the second MEDUSA cycle. We anticipate that comparisons with GHGSat plume detections will be especially important for multispectral instruments, which are challenging to validate with controlled releases due to their higher detection limits. As described in Section 3.3.2, the approach for the acquisition of collocated match-ups between GHGSat and the other missions will be the identification of potential acquisitions by the multispectral and hyperspectral missions in regions where persistent sources and good atmospheric conditions can be expected (e.g. Algeria and Turkmenistan).

We will use the ensemble of high-resolution quantifications as provided by the different groups as baseline for the comparisons.

5.1.4 Task HR-Scientific: Demonstration of the extended validation and intercomparison approach based on plume simulations

During the development of the project, improvements of the current state-of-the-art on plume simulation will be considered in order to recreate conditions closer to real scenes and support the validation exercise as detailed in Section 3.4.2. Improvements that could be considered during these scientific tasks include:

- *Realistic LES simulations*. Setup of conditions that approximate the real scenario on a case-by-case basis (e.g. DEM or land use datasets) (Hald et al., 2019). The simulation of real conditions is expected to provide a larger setup of scenarios and better approximate specific sites.
- *3D end-to-end coupling of LES simulations with Monte-Carlo ray tracing*. The coupling of both emissions and atmosphere in a 3D modelling is expected to improve the understanding of the simulations and validation constraints. The work in Schwaerzel et al. (2020) included 3D simulations of nitrogen dioxide plumes combined with a 3D MonteCarlo ray tracing to study the impact of sun angles in the observations. Work developed as part of the WP4 (see Gorroño et al 2026) simplified this 3D consideration by accounting for the angles during the generation of enhancement maps from 3D simulations. This approach will be considered for the validation and intercomparison with synthetic plumes in this task.

A tentative list of tests to be performed for these demonstrations include the following:

- **Validation 1**: comparison of validation results between the original acquisitions over the controlled release campaigns and those obtained from the same satellite including the simulated plumes (see subsection 3.4.2).
- **Validation 2**: extension to other scenarios that include a) 2-3 other sites with different surface albedo and b) different flux rates.
- **Intercomparison 1**: intercomparison of flux products from EnMAP and Sentinel-2 acquisitions including plume simulations (hyperspectral vs multispectral).
- **Intercomparison 2**: intercomparison of flux products from EnMAP and EMIT acquisitions

including plume simulations (both hyperspectral, but with a different spatial resolution).

The results from this task will be reported either in a stand-alone document or in a dedicated Section of the product validation and intercomparison report. It will summarise the findings of the scientific validation and intercomparison tasks dealing with simulations.

5.2 Tasks for MEDUSA's second validation and intercomparison cycle of methane flux products from global flux mappers

5.2.1 Task FM-1: Intercomparison of flux products from different algorithms (flux mapper)

The intercomparison of detections and flux products from the flux mapper (TROPOMI only in 1st cycle) from different algorithms and groups will be taken from WP300, and extended to more recent data (e.g. all overlapping SRON and Kayrros plumes for 2025) and simulated plumes only focusing on flux rate quantifications (not detections). We will add to that -following same approach- intercomparison also for the locations and dates/times as selected as part of Section 4.3.

To ease the intercomparison of estimates between flux mappers and high-resolution satellites (see following Sections), flux mapper data products will be aggregated together based on their intercomparison results.

5.2.2 Task FM-2: Comparison of flux products from flux mappers with "validated" flux products from high resolution missions

We will intercompare TROPOMI emissions estimates with high-resolution missions validated with controlled-releases, by building a database of match-ups where the source is quite isolated, and the emissions are as stable as possible. A revised version of matchups is provided in Table 4.1.

In order to reduce the number of combinatory pairs of high-resolution estimates (satellites x data providers) comparisons against flux-mapper (satellites x data providers), high-resolution estimates will be aggregated and grouped as described in Section 5.1.1, and flux mapper estimates grouped together as described in Section 5.2.1.

We will also create match-ups to validate the emission estimates from Sentinel-3 SLSTR and GOES-16. As we lack match-ups with high-resolution instruments for these two missions, TROPOMI will serve here as additional reference data in order to get a more significant number of match-ups. See match-ups in Table 4.2 and 4.3.

5.2.3 Task FM-Scientific: Intercomparison of flux mappers with high-resolution instruments using intermediate-resolution instruments as a bridge

As described Section 4.1, we will evaluate whether intermediate resolution instruments (500-750 m; VIIRS, SLSTR) can be used to better understand potential differences in flux rate estimates between flux mappers and high-resolution instruments. As Suomi-NPP VIIRS observes close in time to TROPOMI, plumes can be traced from TROPOMI, through VIIRS and Sentinel-3 overpasses, to Sentinel-2 or other high-resolution instruments. In some cases, the close-proximity overpasses also enable the comparison of retrieved methane enhancements. These intercomparisons could potentially be supported with WRF-LES/WRF simulations, this is to be evaluated during the validation cycle. The results from this task will be reported either in a stand-alone document or in a dedicated Section of the product validation and intercomparison report.

6 Summary

This document describes a series of tasks and tests to be carried out for the validation and intercomparison of satellite-based methane flux products from point sources, with particular application to MEDUSA's second validation and intercomparison cycle of methane flux products. Two interrelated workflows are proposed for high resolution satellites and flux mappers.

In the case of the high-resolution satellites for which the flux products can be directly compared with controlled releases, the latter are taken as the main reference for validation. The approach proposed for flux product validation using concurrent controlled release data relies on the quantification of the regression slope (in line with Sherwin et al. (2023, 2024)) as well as the median relative bias. For the intercomparison between flux products from high resolution systems, we separate between inter-algorithm and inter-satellite comparisons. For inter-algorithm comparisons, we mostly rely on time series of data acquired over controlled-releases, as these can help interpret the differences between the different algorithms; for the inter-satellite comparisons, we rely on existing match-ups between high resolution systems, and are trying to extend this database by requesting more data (in particular from GHGSat). Finally, a scientific task dealing with plume simulations to extend the validation and intercomparison exercise to other sites and flux rate ranges is also proposed.

In the case of flux mappers, their sensitivity to methane and spatial resolution makes it impossible to perform a direct comparison with controlled releases. For this reason, a true validation is not possible, and the assessment of the performance relies on various intercomparisons. This includes intercomparisons of different algorithms for the same instrument, and intercomparisons with high resolution satellites already validated with controlled releases. In a scientific task, using intermediate resolution instruments as a bridge between flux mappes and high-resolution satellites is explored.

The methods described in this protocol will be implemented and tested during MEDUSA's second product validation and intercomparison cycle, and the main results from that exercise will be reported in the update of the PVIR.

7 References

Cusworth, D. H., Jacob, D. J., Varon, D. J., Chan Miller, C., Liu, X., Chance, K., Thorpe, A. K., Duren, R. M., Miller, C. E., Thompson, D. R., Frankenberg, C., Guanter, L., Randles, C. A., 2019. Potential of next-generation imaging spectrometers to detect and quantify methane point sources from space. Atmospheric Measurement Techniques 12 (10), 5655–5668. https://amt.copernicus.org/articles/12/5655/2019/

Cusworth, D. H., Duren, R. M., Thorpe, A. K., Pandey, S., Maasakkers, J. D., Aben, I., Jervis, D., Varon, D. J., Jacob, D. J., Randles, C. A., Gautam, R., Omara, M., Schade, G. W., Dennison, P. E., Frankenberg, C., Gordon, D., Lopinto, E., Miller, C. E., 2021. Multisatellite imaging of a gas well blowout enables quantification of total methane emissions. Geophysical Research Letters 48 (2), e2020GL090864, e2020GL090864 2020GL090864., https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL090864

Dils et al., Product Validation and Intercomparison Report v1, 2026

Foote, M. D., Dennison, P. E., Thorpe, A. K., Thompson, D. R., Jongaramrungruang, S., Frankenberg, C., Joshi, S. C., 2020. Fast and accurate retrieval of methane concentration from imaging spectrometer data using sparsity prior. IEEE Transactions on Geoscience and Remote Sensing 58 (9), 6480–6492. https://doi.org/10.1109/TGRS.2020.2976888

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. https://doi.org/10.5194/amt-16-89-2023

Gorroño, J., Pei, Z., Valverde, A., and Guanter, L.: Considering the observation and illumination angular configuration for an improved detection and quantification of methane emissions, Atmos. Meas. Tech., 19, 1245–1257, https://doi.org/10.5194/amt-19-1245-2026, 2026.

Gorroño, J., Pei, Z., & Guanter, L. (2026). Benchmark simulations for methane emissions validation and sensitivity studies [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18161182

Groshenry, Alexis & Giron, C. & Hessel, C. & Franchis, C. & Facciolo, Gabriele & Ehret, T.. (2024). Methane Emissions Monitoring Using Geostationary Satellites. 10376-10380. 10.1109/IGARSS53475.2024.10642856.

Guanter, L., Irakulis-Loitxate, I., Gorroño, J., Sa´nchez-García, E., Cusworth, D. H., Varon, D. J., Cogliati, S., Colombo, R., 2021. Mapping methane point emissions with the prisma spaceborne imaging spectrometer. Remote Sensing of Environment 265, 112671. https://www.sciencedirect.com/science/article/pii/S0034425721003916

Guanter L., 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
Environmental Science & Technology Letters 2024 11 (8), 825-830, https://doi.org/10.1021/acs.estlett.4c00399

Guanter, L., Warren, J., Omara, M., Chulakadabba, A., Roger, J., Sargent, M., Franklin, J. E., Wofsy, S. C., and Gautam, R.: Detection and quantification of methane plumes with the MethaneAIR airborne spectrometer, Atmos. Meas. Tech., 18, 3857–3872, https://doi.org/10.5194/amt-18-3857-2025, 2025.

Hald, C., Zeeman, M., Laux, P., Mauder, M., Kunstmann, H., 2019. Large-eddy simulations of real-world episodes in complex terrain based on era-reanalysis and validated by ground-based remote sensing data. Monthly Weather Review 147 (12), 4325 – 4343. https://journals.ametsoc.org/view/journals/mwre/147/12/mwr-d-19-0016.1.xml

He, T.L., Varon, D.J., Kondragunta, S., Ren, X., Cohen, M., Carroll, B.J., Malarich, N., Peischl, J., de Jong, T.A., Gerritsen, J. and Maasakkers, J.D., 2026. Coordinated satellite, aircraft, and ground-based observations of a large transient methane release. https://doi.org/10.31223/X54R07

Irakulis-Loitxate, I., Guanter, L., Maasakkers, J. D., Zavala-Araiza, D., Aben, I., 2022. Satellites detect abaTable super-emissions in one of the world's largest methane hotspot regions. Environmental Science & Technology 56 (4), 2143–2152, pMID: 35102741. https://doi.org/10.1021/acs.est.1c04873

De Jong, T. A., Maasakkers, J. D., Irakulis-Loitxate, I., Randles, C. A., Tol, P., & Aben, I. (2025). Daily global methane super-emitter detection and source identification with sub-daily tracking. Geophysical Research Letters, 52(8), e2024GL111824. https://doi.org/10.1029/2024GL111824

He, T.L., Varon, D.J., Kondragunta, S., Ren, X., Cohen, M., Carroll, B.J., Malarich, N., Peischl, J., de Jong, T.A., Gerritsen, J. and Maasakkers, J.D., 2026. Coordinated satellite, aircraft, and ground-based observations of a large transient methane release, https://doi.org/10.31223/X54R07

Jongaramrungruang, S., Matheou, G., Thorpe, A. K., Zeng, Z.-C., Frankenberg, C., 2021. Re- mote sensing of methane plumes: instrument tradeoff analysis for detecting and quantifying local sources at global scale. Atmospheric Measurement Techniques 14 (12), 7999–8017, https://amt.copernicus.org/articles/14/7999/2021/

McKeever, J. and Jervis, D.: Validation and Metrics for Emissions Detection by Satellite, Tech. rep., GHGSat, https://www.ghgsat.com/en/scientific-publications/validation-and-metrics-for-emissions-detection-by-satellite/

Ouerghi, E., Ehret, T., Facciolo, G., Meinhardt, E., Marion, R., and Morel, J.-M.: Tightening up methane plume source rate estimation in EnMAP and PRISMA images, Atmos. Meas. Tech., 18, 4611–4629, https://doi.org/10.5194/amt-18-4611-2025, 2025.

Pandey, S., van Nistelrooij, M., Maasakkers, J. D., Sutar, P., Houweling, S., Varon, D. J., Tol, P., Gains, D., Worden, J., Aben, I., 2023. Daily detection and quantification of methane leaks using sentinel-3: a tiered satellite observation approach with sentinel-2 and sentinel-5p. Remote Sensing of Environment 296, 113716., https://www.sciencedirect.com/science/article/pii/S0034425723002675

Frances E. Reuland, Taylor Adams, Evan D. Sherwin, Sahar H. El Abbadi, Eric A. Kort, Adam R. Brandt, Unlocking credible space-based methane sensing through a year-long single-blind test, 2026, PREPRINT (Version 1), https://doi.org/10.21203/rs.3.rs-9110475/v1

Roger, J., Irakulis-Loitxate, I., Valverde, A., Gorroño, J., Chabrillat, S., Brell, M., Guanter, L., 2024. High-resolution methane mapping with the EnMAP satellite imaging spectroscopy mis- sion. IEEE Transactions on Geoscience and Remote Sensing, 1–1. https://doi.org/ 10.1109/TGRS.2024.3352403

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.

Schwaerzel, M., Emde, C., Brunner, D., Morales, R., Wagner, T., Berne, A., Buchmann, B., Kuhlmann, G., 2020. Three-dimensional radiative transfer effects on airborne and ground- based trace gas remote sensing. Atmospheric Measurement Techniques 13 (8), 4277–4293. https://amt.copernicus.org/articles/13/4277/2020/

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, 3836. https://doi.org/10.1038/s41598-023-30761-2

Sherwin, E. D., El Abbadi, S. H., Burdeau, P. M., Zhang, Z., Chen, Z., Rutherford, J. S., Chen, Y., Brandt, A. R., 2024. Single-blind test of nine methane-sensing satellite systems from three continents. Atmospheric Measurement Techniques 17 (2), 765–782. https://amt.copernicus.org/articles/17/765/2024/

Watine-Guiu, M., Varon, D. J., Irakulis-Loitxate, I., Balasus, N., Jacob, D. J., 2023. Geostationary satellite observations of extreme and transient methane emissions from oil and gas infrastruc- ture. Proceedings of the National Academy of Sciences 120 (52), e2310797120., https://www.pnas.org/doi/abs/10.1073/pnas.2310797120

Yu, Q., Jervis, D., Huang, Y., 2024. Accounting for the effect of aerosols in ghgsat methane retrieval. Atmospheric Measurement Techniques 17 (11), 3347–3366.
URL https://amt.copernicus.org/articles/17/3347/2024/

Appendix A On the importance of acquisition time difference for the intercomparison of flux products from different satellites

The intercomparison of methane flux products from different satellite missions is challenging because sources can be highly variable and methane can be rapidly transported in the atmosphere. Therefore, it is possible that the comparison between missions represents different plume morphologies within minutes and different flux rates and scene conditions (e.g. wind profiles) for longer timespans. Here, we evaluate the impact on flux product intercomparisons of time differences in data acquisition.

The IMEO-MARS database contains thousands of flux rates between missions and, we can directly compare the flux rates from different satellites and filter the results based on spatiotemporal distance. We have selected the Copernicus Sentinel 2 and Landsat flux rates and applied different filtering steps that include:

- Matching sources are considered within 50m distance
- Wind differences must be below 0.2m/s. Rarely occur within minutes (for hourly-integrated products) but filters very important wind variations.
- Flux rates below 10t/h. We assume that above this flux rate, many events could be short-term unstable releases including explosions.
- Differences above 80% are also filtered assuming that they correspond to outliers. This is a conservative number that represents a combined uncertainty with approximately 60% uncertainty for each flux rate.

Figure A1 presents the scatter plot for the flux rate comparison between the Copernicus Sentinel 2 and Landsat mission applying the filtering steps described above and a time difference of up to 15 minutes.

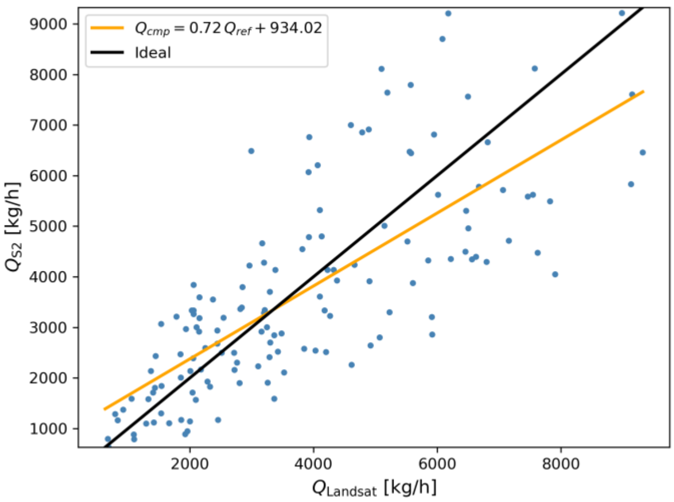


Figure A1: Scatter plot of flux rates between Landsat and Copernicus Sentinel 2 for temporal distances up to 15 minutes using the IMEO database. The comparison looks for matches within 50 m distance, below 0.2 m/s wind differences and flux rates below 10t/h. Differences above 80%

are also filtered assuming that they correspond to outliers.

The results in Figure A1 show a total of 137 matches with considerable correlation of 0.75 but an important intercept increase with flux rate values. The RMSE and absolute errors are between 1000-1500kg/h with a mean relative error of 33%.

We repeat the same exercise with variable time differences sweeping timespans that range from 20s up to 15 minutes. Figure A2 illustrates the evolution of slope and intercept values, while Figure A3 includes the correlation and mean relative error.

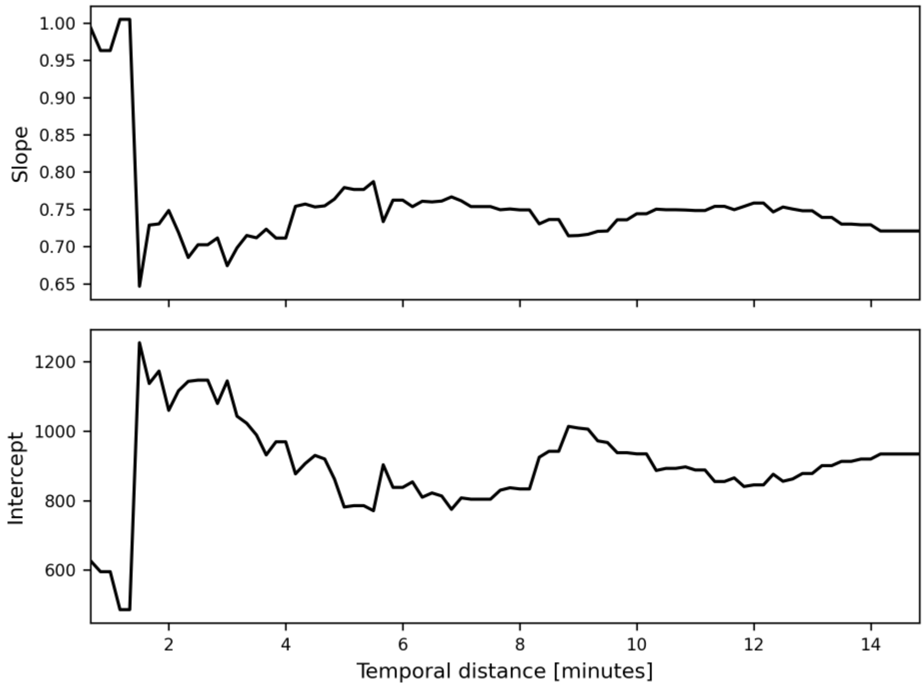


Figure A2: Temporal trend of slope and intercept for comparisons between Landsat and Copernicus Sentinel 2 using the IMEO database. The comparison looks for matches within 50m distance, below 0.2m/s wind differences and flux rates below 10t/h. Differences above 80% are also filtered assuming that they correspond to outliers.

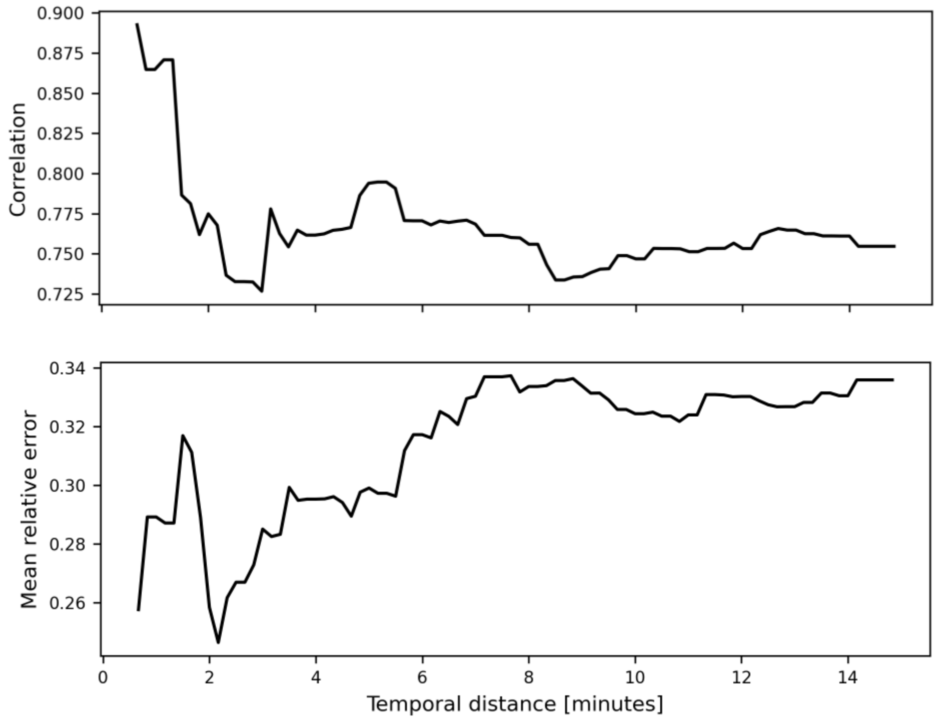


Figure A3: Temporal trend of correlation and mean relative error for comparisons between Landsat and Copernicus Sentinel 2 using the IMEO database. The comparison looks for matches within 50m distance, below 0.2m/s wind differences and flux rates below 10t/h. Differences above 80% are also filtered assuming that they correspond to outliers.

The results show that, over the first minute, the correlation is higher, the slope is almost ideal while the intercept decreases to 600kg/h. The parameters rapidly change after the first two minutes with a similar pattern as those presented in Figure A1 after 15 minutes. For the first two minutes, the number of matches ranges from 10 to 20 and might not be statistically robust. However, the results suggest a similar pattern as described in Ouerghi et al. (2025) where the plumes are highly spatially correlated for time differences below 2 minutes. In short, we can separate three different intercomparisons based on time difference (the time ranges are approximate):

- **0-2 minutes**. Pseudo-invariant ΔXCH4 and Q. The plumes exhibit similar spatial patterns, and the comparison highlights biases and offsets in the retrieval and quantification.
- **2-15 minutes**. Pseudo-invariant Q. The spatial pattern is significantly different. The flux rate is expected not to significantly change but the plume morphology does. The intercomparison error is expected to include other effects such as the calibration fitting dispersion or sensitivity to scene surface variations.
- **15 minutes and over**. Highly variable ΔXCH4 and Q. The intercomparison can be largely affected by scene variations. These include different wind profiles, turbulences, and illumination patterns. More importantly, depending on the specific source it might be possible that there are changes in the emitted flux rate.