NIST Interagency Report
NIST IR 8575

Common Practices for Quantifying Methane Emissions from Plumes Detected by Remote Sensing

A collaborative report of CEOS, NPL, NIST, and LBNL

John Worden
Paul Green
Annmarie Eldering
Evan Sherwin

NIST | NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Common Practices for Quantifying Methane Emissions from Plumes Detected by Remote Sensing

Collaborative report of CEOS, NPL, NIST, and LBNL

John Worden
National Aeronautics and Space Administration (NASA)

Annmarie Eldering
Special Programs Office
Laboratory Programs, NIST

Paul Green
Earth Observations & Climate Group
National Physics Laboratory, UK
CEOS GHG Task Team

Evan Sherwin
Energy Analysis & Env. Impacts Division
Lawrence Berkeley National Laboratory

May 2025

NIST IR 8575
May 2025

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

This publication is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and other organizations. This report was prepared as an account of a workshop; it is intended to document external perspectives and does not represent official NIST positions.

Author ORCID iDs
John Worden: 0000-0003-0257-9549
Paul Green: 0000-0003-1543-1658
Annmarie Eldering: 0000-0003-1080-9922
Evan Sherwin: 0000-0003-2180-4297

Contact Information
annmarie.eldering@nist.gov

Abstract

This document provides a set of community-accepted practices for quantifying methane emissions based on plumes detected via spectroscopic remote sensing. Its primary goal is to promote consistency in the generation, validation, reporting, and quality assessment of methane emission estimates derived from remote sensing radiances. Developed by subject matter experts with deep experience across all stages of the measurement process, this guidance reflects a critical evaluation of current methodologies and highlights key practices needed to produce reliable, interoperable, and traceable products. The focus is specifically on methane emissions quantified from distinct plumes originating from localized sources, rather than diffuse emissions spread over large regions, which are beyond the scope of this work. This document is intended to serve both data producers and users. For producers, it offers a framework for aligning with field-recognized standards to ensure their outputs meet rigorous quality and transparency criteria. For users, it provides a reference to assess dataset fitness-for-purpose by highlighting essential metadata, assumptions, and methodological choices that underpin emission estimates. By fostering a shared understanding of best practices, this work aims to enhance comparability, confidence, and utility of remotely sensed methane emission products.

Keywords

Table of Contents

List of Tables

List of Figures

Executive Summary

This document provides a set of community-accepted practices for quantifying methane emissions based on plumes detected via spectroscopic remote sensing. Its primary goal is to promote consistency in the generation, validation, reporting, and quality assessment of methane emission estimates derived from remote sensing radiances. Developed by subject matter experts with deep experience across all stages of the measurement process, this guidance reflects a critical evaluation of current methodologies and highlights key practices needed to produce reliable, interoperable, and traceable products. The focus is specifically on methane emissions quantified from distinct plumes originating from localized sources, rather than diffuse emissions spread over large regions, which are beyond the scope of this work. This document is intended to serve both data producers and users. For producers, it offers a framework for aligning with field-recognized standards to ensure their outputs meet rigorous quality and transparency criteria. For users, it provides a reference to assess dataset fitness-for-purpose by highlighting essential metadata, assumptions, and methodological choices that underpin emission estimates. By fostering a shared understanding of best practices, this work aims to enhance comparability, confidence, and utility of remotely sensed methane emission products.

The document is organized as follows:

- **Section 1** presents the motivation for this document and the growing need for consistent, transparent practices to support the quantification of methane emissions based on observed plumes using spectroscopic remote sensing. It outlines the timeliness of articulating community-accepted guidance as plume-based approaches become increasingly central to emissions monitoring and decision-making.

- **Section 2** describes the current, community-accepted methods used to derive methane emission estimates from plume observations based on spectroscopic remote sensing. It outlines typical workflows for plume detection and quantification, highlighting the methodological components required to ensure transparent, reproducible, and interoperable emission products.

- **Section 3** outlines the commonly applied approaches for validating plume detections and associated emission estimates. It emphasizes the importance of using independent datasets and clearly defined metrics to evaluate the accuracy and reliability of plume-based remote sensing observations.

- **Section 4** provides a framework for assessing the quality of retrieved methane column values, such as enhancements relative to background or total column amounts. This section identifies the metadata and quality indicators necessary to interpret column observations and ensure they support robust plume-based emission quantification.

Contributors

Section 2 contributing authors:

- Dan Cusworth, Carbon Mapper (key contributor to structure of the section)
- Jason McKeever, GHGSat
- Dylan Jervis, GHGSat
- Daniel Varon, Harvard
- Andrew Thorpe, NASA JPL, EMIT team
- Philip Brodrick, NASA JPL, EMIT team
- James Whetstone, NIST
- Riley Duren, Carbon Mapper
- Adam Brandt, Stanford University

Section 3 contributing authors:

- Evan D. Sherwin, Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area and Energy Technologies Area
- Adam R. Brandt, Stanford University, Department of Energy Science & Engineering

Section 4 contributing authors:

- John Worden, NASA JPL
- Samuel E. Hunt, NPL
- Patrick Barker, NPL
- Paul Green, NPL
- Jean-Christopher Lambert, BIRA-IASB
- Jaime Nickeson, SSAI/NASA GSFC
- Stephen Compernolle, BIRA-IASB
- Benjamin Poulter, NASA
- Angelika Dehn, ESA
- Sabrina Pinori, Serco
- Lidia Saavedra De Miguel, ESA
- Amy Beaton, Telespazio
- Gabriele Mevi, Serco
- Kevin Halsall, Telespazio

1. Background

Collaborations between industry and academia—such as the Energy Emissions Modeling Data Laboratory (EEMDL)—are advancing our ability to characterize and quantify methane emissions through new data sources and modeling approaches. These initiatives are complemented by the rapid development and deployment of satellite-based methane monitoring technologies (Jacob et al., 2022), which improve the accuracy, resolution, and frequency of emissions detection.

The pursuit of increased productivity alongside emissions reductions (Lu et al., 2023) has underscored the need for trusted, independently verifiable emissions data. Satellite-derived measurements are central to this effort. The global reach and inherent spatial sampling capabilities of on-orbit instruments make them uniquely suited for consistent, repeatable surveys across regions and borders. These systems enable the detection and quantification of methane plumes, the identification of major sources and sinks, and the attribution of emissions to specific geographic locations.

This document focuses specifically on the quantification of emissions derived from observed methane plumes. These measurements have primarily been applied to the fossil energy and waste sectors (Cusworth et al., 2022; Thorpe et al., 2023), with current satellites typically detecting emissions exceeding $\approx$100 kg CH_4 per hour, while airborne platforms can observe sources as small as $\approx$10 kg CH_4 per hour.

A growing number of on-orbit sensors—launched by commercial, philanthropic, and other non-governmental organizations—now complement long-standing public-sector missions. These sensors observe methane plumes under favorable conditions and estimate emissions from known sources. Additionally, several legacy satellite platforms not originally intended for methane detection have demonstrated the capability to detect large plumes, broadening the range of usable assets. This has enabled a diverse group of stakeholders—including academia, space agencies, start-ups, and international organizations—to contribute to the methane monitoring ecosystem.

Plume-based observations are particularly valuable for rapidly identifying and quantifying the largest sources of emissions, enabling timely mitigation. While the share of total anthropogenic methane emissions attributable to such sources remains uncertain, regional estimates indicate they can be significant. For example, in the Permian Basin, sources exceeding 100 kg CH_4/hour were estimated to contribute 75% of total oil and gas emissions in 2019 (Sherwin et al., 2024b), with other analyses suggesting a lower share of 23% in 2021 (Williams et al., 2025). In basins such as the Uinta and San Joaquin, the corresponding contributions range from 10% to 50%. Outside these intensively studied regions, the observable fraction remains poorly constrained. Even at the lower end of these estimates, plume-based observations represent a critical tool for emissions mitigation. Broader assessments of sector-wide or regional emissions, particularly from diffuse sources such as agriculture, will require complementary approaches like area-based flux mapping using instruments such as MethaneSAT and TROPOMI.

Despite the rapid growth in observational capacity, challenges remain. Divergent emissions estimates, opaque methodologies, and inconsistent validation approaches can erode

confidence in remote sensing-based emissions data. The emergence of non-public-sector missions using proprietary methods—often without full transparency across the data chain— further highlights the need for community-accepted practices to ensure traceability, comparability, and scientific credibility.

To address this need, the greenhouse gas (GHG) community—through the Committee on Earth Observation Satellites (CEOS) and National Metrology Institutes (NMIs)—has developed this document to articulate commonly accepted approaches for quantifying methane emissions based on observed plumes. It provides guidance spanning from Level 0/1 radiance, to Level 2 concentration, to Level 4 emissions, and includes current practices for validation and quality assessment. The focus is on emissions derived from discrete plumes, rather than from spatially diffuse sources.

The practices articulated here aim to increase transparency and trust in satellite-based emissions products, especially as the field grows to include a diverse and expanding set of data producers. Looking forward, this document may also serve as a foundation for the development of a future community framework for satellite-based area flux estimation, which is needed to quantify emissions from diffuse sources that fall outside the scope of plume detection.

2. Common Practices for the Identification of Methane Plumes and Quantifying Emissions at the Facility Scale

2.1. Background

This section captures the current state of implementation and common practices in the analysis of remote sensing data for methane plume detection, with a focus on facility-scale emissions of hundreds of kilograms per hour and larger. The rapid evolution of the field, frequent addition of new measurement instrumentation and analysis teams, and the wide range of capabilities and results points to the need for an effort to work towards consensus standards. A first step is to document and discuss the analysis processes across teams, with a focus on developing a common taxonomy and understanding potential reasons for variability of results across teams. An important outcome of collecting these common practices is the establishment of a taxonomy. Thus, we document agreed-upon definitions and link them to reference documents, such as the Joint Committee for Guides in Metrology's Guide to the Expression of Uncertainty in Measurement (JCGM GUM), and the International Vocabulary of Metrology (JCGM VIM) published by the BIPM.

The general analysis steps of remote sensing data for methane plume detection are outlined in Figure 1 and are discussed briefly in this section. In the sections that follow, each step is discussed in more detail, as well as definitions and open issues. In the future, intercomparison of teams results and analysis of controlled release measurement datasets will be performed to add to insights about variability across teams and the impact of analysis choices. This is the first effort to capture and integrate information, so some generalizations are made and not all variations of approaches are included here. We will note where simplifications are made or additional information is expected to be included in future updates.

The instruments discussed here are spectrally resolved imagers that gather measurements of reflected sunlight in the spectral regions where methane absorbs light (around 1.6 μm and/or 2.3 μm). These data are collected using 2D sensors, resulting in image-like maps with a third dimension representing the wavelength of light. The radiance measurements are made with spectral resolutions ranging from 0.3 nm to > 100 nm. Several papers document the range of instrumentation and missions, such as Cusworth et al. (2019), Jacob et al. (2022), and Jongaramrungruang et al. (2021). These instruments range from systems specifically designed to measure methane plumes to instruments designed for other purposes (such as global, large scale methane characterization or imagery collection) that have been found to have sensitivity to methane and can be used to define and quantify plumes. The wide range of capabilities and design (spectral resolution, spatial resolution, area measured, etc.) also contribute to a wide range of analysis approaches and strategies for characterizing plumes, thus contributing to the motivation for development of consensus standards.

Radiance measurements are then used to estimate methane concentrations or concentration enhancements in the area of interest. This is generally done through a physics-based retrieval approach (e.g., Iterative Maximum a Posteriori - Differential Optical Absorption Spectroscopy, or IMAP-DOAS) or a statistical method, such as a matched filter (MF) for imaging spectrometers

(Thorpe et al. 2023), and with multi-temporal composites of band ratios for multi-spectral imagers (e.g., Varon et al., 2021). The next step is to identify methane enhancements, or plumes, within the data field. Various approaches are used for plume detection, ranging from manual identification by experts (Varon et al., 2021) to automated methods, including machine learning, which identifies pixels with higher methane concentrations than the background (Redout-Leduc et al., 2024).

Once the plume is identified, the emission rate can be estimated. Several methods are used, with the integrated mass enhancement (IME) method being one of the most frequently applied (Frankenberg et al., 2016; Varon et al., 2018; Duren et al., 2019; Jongaramrungruang et al., 2019; Jacob et al., 2022). As implemented by Varon et al. (2018), the IME method calculates the source rate (Q) using the total plume IME (kg), an effective wind speed (U_{eff}, m·s^{-1}), and a plume length scale (L, m).

Key ancillary data, including wind speed and atmospheric mixing (diffusion rates), are not directly measured by the plume mapper instruments and must be estimated from other data sources. Wind direction may be taken from external data sources or estimated from the measurement data. These factors significantly impact methane flux estimates (Sherwin et al., 2023, 2024a). Different groups may use different wind data sources and make different assumptions about atmospheric mixing, which are discussed later in this section.

For further insight into the measurement concepts and additional background information, readers can search for resources such as those provided by Carbon Mapper in FAQs, Bridger Photonics Methane Dectection, and Kayrros Technology Overview. Other sites include:

- UNEP International Methane Emissions Observatory
- NASA Methane Source Finder

2.2. Processing of Satellite Observations from L0 (raw observation) to L4 (Emissions)

Figure 1 captures the typical analysis steps in the methane plume identification and quantification process. The inputs and output data for each step are annotated. In the following subsections, each step is described in detail, including definitions and current practices. We also note some outstanding issues and unresolved questions throughout.

Key products/metrics that can be
derived from various stages of workflow
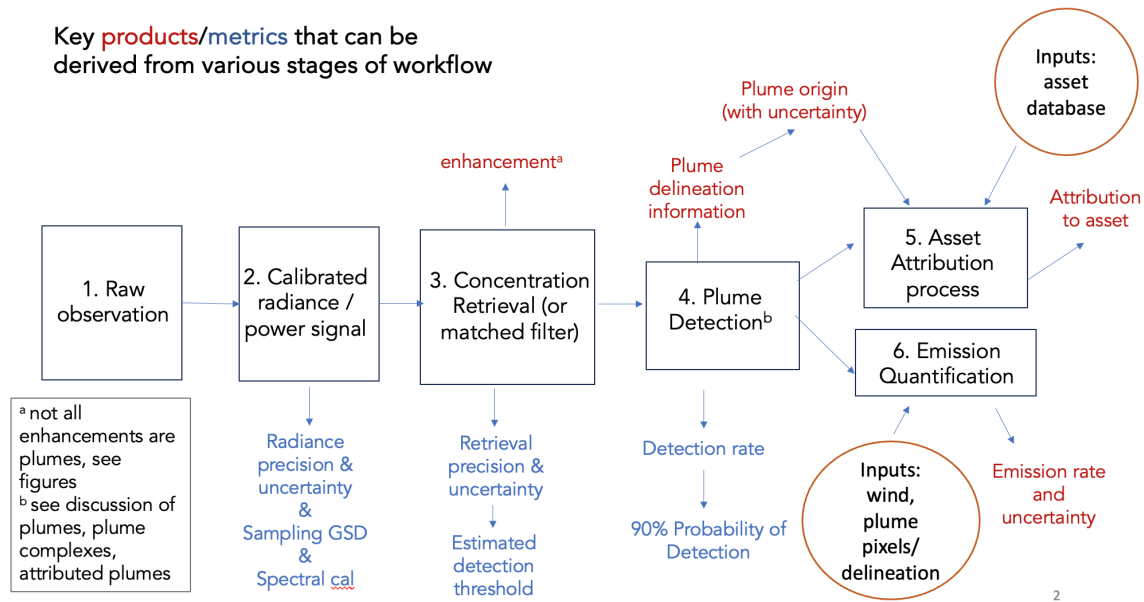
Figure 1. Typical Analysis Steps for Methane Plume Detection and Quantification Process. Output products are in brick colored font, while useful metrics from each step are in blue. The background section includes a high-level description of these steps, and details are included in sections that follow. Concept credit: Dan Cusworth, Carbon Mapper, used with permission.

2.3. Calibrated Radiance/Power Signal Element
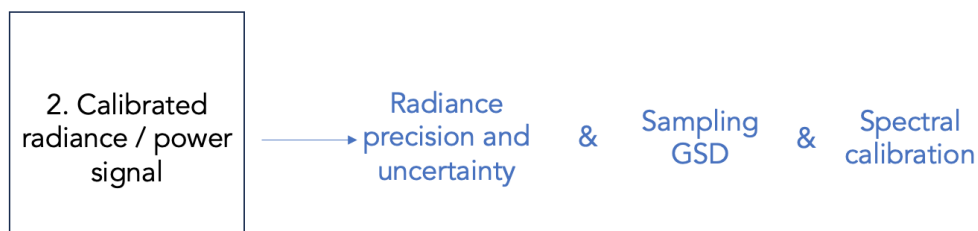
Figure 2. Calibrated Radiance and Key Output Metrics. Concept credit: Dan Cusworth, Carbon Mapper.

This section outlines three key considerations when working with calibrated radiances. Understanding and documenting the radiance precision, spectral calibration, and ground sampling distance (GSD) parameters in methane plume detection workflows parameters is crucial for accurate analysis and interpretation of remote sensing data, especially when identifying and quantifying methane emissions.

2.3.1. Calibrated Radiance

The calibrated radiance serves as the starting point for many teams. Raw observations collected by the instrument are transformed to calibrated radiance based on pre-flight instrument characterization. The preferred SI unit for radiance data is $W/(m^2 \cdot \mu m \cdot sr)$. Several important characteristics should be recorded alongside the radiance, including the spectral grid and

details about spectral sampling, such as the instrument line shape (ILS) or the full width at half maximum (FWHM) of the spectral response function. Additionally, it is necessary to record the signal-to-noise ratio (SNR) of the radiance, or a measure of noise as a function of wavelength. Lastly, information about the instrument's spatial response is required.

2.3.2. Radiance Uncertainty

Radiance uncertainty arises from random and systematic sources of error. The radiance precision (random error) is driven by several factors, which are determined by the instrument's characteristics, including detector noise, detector efficiency, transmission efficiency, signal level, and integration time. Systematic errors may be related to offsets in electronics, error in calibration, systematic changes in the instrument, unaccounted for temperature sensitivity, and other factors. Pre-flight calibration and characterization provide an initial estimate of uncertainty before launch, and on-board methods allow for monitoring and updates throughout the mission's lifetime. Pre-flight radiometric calibration typically involves the use of reference standards, such as lamp- and laser-illuminated integrating spheres, traceable to the International System of Units (SI) via a National Metrology Institute (NMI).

The radiometric requirements for plume detection are generally less stringent than those for measuring variations in background methane concentration fields, although instruments with lower spectral resolution (less information on spectral lines) may have more sensitivity to radiometric calibration and instrument noise characteristics (Jongaramrungruang et al., 2021). Typically, emphasis is placed on the linearity of the measurement system and the radiance precision. The plume detection approach is focused on enhancements above background concentrations, so it is believed to be less impacted by miscalibration of absolute radiance than other approaches, although nonlinearity, additive offsets, or spurious radiance offsets are of concern. Some in-flight radiometric verification can be conducted addressing both the radiance precision and uncertainty, using instrumented sites like those provided by RadCalNet (Bouvet et al., 2019).

2.3.3. Ground Sampling Distance and Point Spread Function

GSD is defined as the distance between the centers of two adjacent samples or pixels on the ground while the point spread function (PSF), an array of information about the spatial distribution of light that is seen by each pixel, is an integrated description of the performance of the full imaging system. These characteristics vary with parameters such as orbit altitude, the off-nadir angle, cross-track and along-track instrument characteristics, and other factors.

The GSD and the PSF are generally characterized pre-launch in the laboratory. This information must be combined with the location of the sensor (e.g., the position along an orbit track of a satellite) and its viewing direction to geo-locate each footprint relative to a surface map. In-flight verification can be performed using ground features such as coastlines, bridges, or small, isolated landmarks to assess the spatial sampling performance. Understanding of GSD and PSF is crucial when determining the location of emission sources, as it significantly contributes to

the uncertainty in source location. Source location data should always include the GSD and/or PSF characteristics, as they directly influence the precision of source geolocation.

2.3.4. Spectral Calibration

The characterization of the spectral calibration of the instrument is also important and impacts the next step of matched filter analysis. The spectral calibration typically includes measurement of the spectral response functions (SRF) as well as assessment of spectral calibration, or assignment of the central wavelengths to each pixel. In the retrieval steps that follow, models of the atmospheric absorption are convolved with the SRF to create simulated data that can be compared to the instrument measurement data.

2.4. Concentration Retrieval/Matched Filter
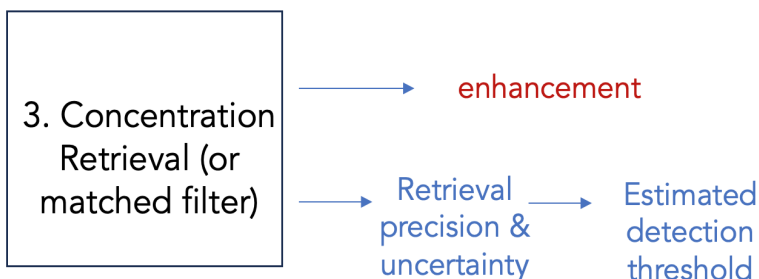
Figure 3. Concentration Step Products and Metrics. Concept credit: Dan Cusworth, Carbon Mapper, used with permission.

This section provides an overview of current approaches used for concentration and enhancement retrievals (Figure 3) in methane detection workflows, including the range of techniques. We also highlight the importance of retrieval uncertainty and common practices related to quantifying concentration enhancements. Figure 3 illustrates the concentration retrieval and enhancement step within the broader methane detection workflow.

2.4.1. Concentration Retrievals

There are several different approaches currently in use for concentration (enhancement) retrievals. Some examples of techniques include those cited below, but approaches are evolving rapidly, so this list may become out of date. Ideally, the concentration retrieval process also provides an estimate of retrieval uncertainty to ensure robust analysis.

- **Matched Filter** (e.g., Thompson et al., 2015; Foote et al., 2020)

- **Band Difference/Ratio** (e.g., Varon et al., 2021)

- **IMAP-DOAS** (typically assumes no scattering) (Frankenberg et al., 2005; Frankenberg et al., 2016; Thorpe et al., 2017)

- **Full Physics** (including scattering) (Parker et al., 2011; O'Dell et al., 2012 [for CO_2])

2.4.2. Retrieval Uncertainty

As with radiance uncertainty, retrieval uncertainty arises from random and systematic sources of error. No details on the characterization of systematic error have been captured in this report. Retrieval uncertainty is critical because, during the process of identifying enhancements relative to the background, both the uncertainty and the concentration resolution granularity will influence the results. Generally, higher retrieval uncertainty limits the ability to identify smaller enhancements.

Two working definitions of retrieval precision (random error) are proposed in this consensus standards work. Figure 4 shows an example of calculating variability in the background region, highlighting two approaches to handling retrieval uncertainty: empirical methods and posterior error covariance. The working definitions are outlined below:

- Bayesian Optimal Estimation (OE) Retrieval Precision: This approach uses the posterior error covariance from an optimal estimation retrieval. Some additional considerations are:

  - In OE retrievals, constraints, priors, and other assumptions can significantly affect the retrieval uncertainty. In general, very large prior covariances have less of an impact on the posterior retrieval output. A discussion of the impact of priors and prior misspecification can be found in Nguyen et al. (2019).

  - Precision of column abundance retrieval is typically predicted from theory, based on the amount of collected light (shot noise) and camera specifications (readout noise), or is estimated from fit residuals.

  - While useful for design and analysis, these approaches are thought to underestimate the impact of artifacts and unmodeled physical effects.
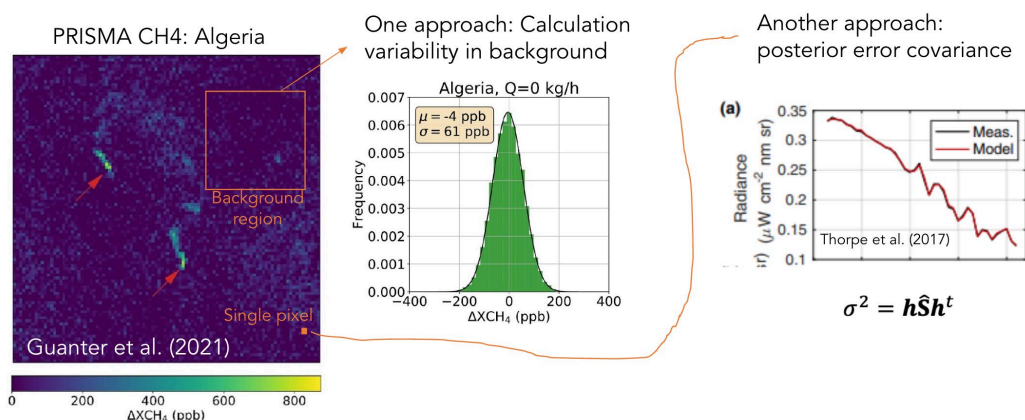
Figure 4. Concentration Precision Analysis Approaches. Credit: Dan Cusworth, Carbon Mapper, used with permission.

- Empirical (Background) Retrieval Precision: This precision is empirically estimated based on column retrievals, obtained through replicate measurements on the same or similar objects under specified conditions, typically using background measurements where no methane plumes are present. Additional considerations are:

- o A practical approach is to calculate the spatial standard deviation within a region of interest in the retrieval domain, where there are no methane emissions.

- o When assessing methane enhancements above background levels (delta XCH4 on Figure 4), the mean value should be zero, and any variability reflects the uncertainty of the retrieved methane.

- o Observing conditions must also be documented. Ideally, empirical measurement precision should be calculated under the same albedo, solar zenith angle (SZA), and viewing angle, so the results are comparable. Because this condition is rarely met, any reported precision must include the viewing conditions.

2.4.3. Concentration Enhancement

In practice, concentration enhancement refers to the analysis step where the background concentration or enhancement field is defined, and pixels with concentrations elevated above this background are identified. In the case of matched filters, the background and pixels with enhanced concentration are defined, but with the field of enhancement data in the total column ($\mu mol \cdot mol^{-1} \cdot m$ or ppm $\cdot$m), rather than concentration (Foote et al., 2020). This is distinct from plume detection in that there may be connected pixels representing an enhancement, or scattered pixels of enhancement. Enhanced pixels can sometimes follow land features or roads due to errors correlated with surface reflectance. The enhancement step of the workflow is a general concept that identifies pixels with elevated concentrations, which then feeds into the subsequent plume detection step, discussed later.

2.5. Plume detection
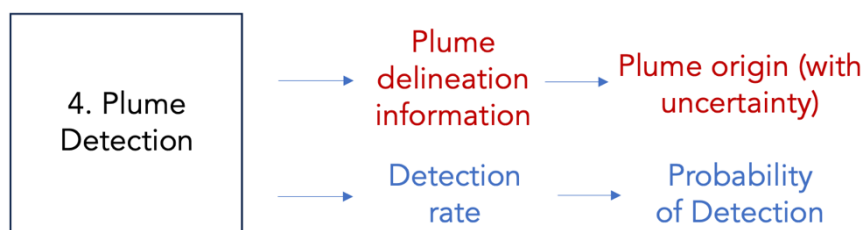
Figure 5. Plume Detection Step with Outputs and Key Metrics. Concept credit: Dan Cusworth, Carbon Mapper.

Example of plume segmentation process
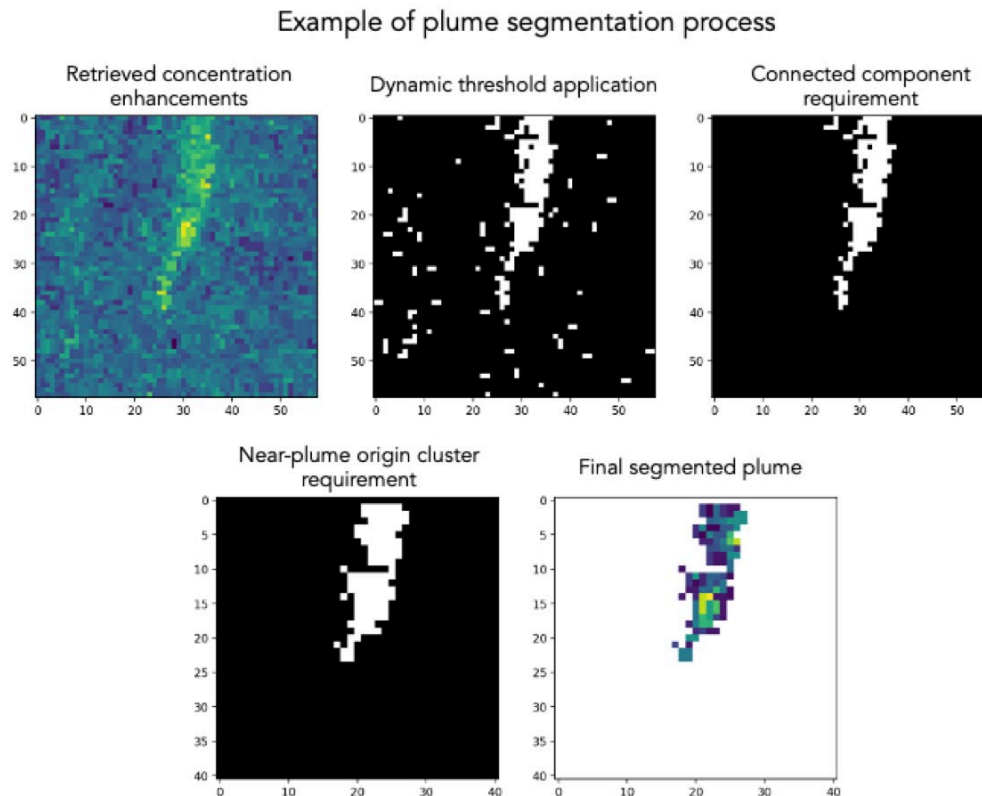
Figure 6. Steps Showing the Process of Finding the Methane Enhancement and Then the Plume. Credit: Dan Cusworth, Carbon Mapper, used with permission.

Plume segmentation is a critical step (Figure 5) in methane plume detection through remote sensing. This process involves selecting and grouping pixels that show enhanced methane concentration levels to define the spatial boundaries of the plume (Figure 6). Currently, there is no comprehensive review available that details the methods for plume segmentation. In many cases, the segmentation approach is presented alongside the quantification methods in scientific publications. However, this is an area of rapid technological advancement. Plume segmentation provides essential parameters that are needed for accurate emission quantification, such as the plume characteristic length (L). While the methods for plume segmentation are still developing, the process is integral to the accurate estimation of methane emissions. This section reviews the common practices and techniques used in plume segmentation and discusses the challenges that remain in standardizing this crucial step in light of the new emerging techniques.

Common approaches include:

- Clumping algorithms
- Machine learning
- Visual analysis or hand-drawn methods

- Percentile thresholding

It is common to have a manual or human review of plume identification as part of the quality control before the next steps of plume quantification. While the manual review can in ways result in strict quality controls, it also may introduce bias and repeatability challenges, which could affect the overall detection and quantification statistics and performance characteristics. As a result, further development is needed in the best practices to mitigate this bias and introduce automation as the volume of measurements rapidly increases.

Key considerations for improving consistency and reducing operator impact include:

- Using objective measures such as signal-to-error ratio

- Providing an error estimate for each pixel (empirical value) that is an input to plume segmentation

- Accounting for scene artifacts, where possible, and develop automated differentiation between artifacts and true enhancements

2.5.1. Discussion of Current Practices

The consensus from the community is that plume segmentation approaches have many common elements. Typically, the enhanced region is separated from the background based on signal levels above noise or using thresholding techniques. However, different use cases may lead to varying thresholding approaches. These are outlined below:

- **Case 1: Visualization:** Plume segmentation for visualization purposes, used by some groups, often uses lower thresholds, which results in larger plume extents. This approach is mainly used for communication and detection purposes.

- **Case 2: Emissions Quantification:** For emissions quantification, a higher threshold is applied, leading to a more restricted plume extent.

- **Case 3: Public Hazard Notification:** In cases where concentration enhancements are used to inform hazard notifications, conservative plume segmentation may be practiced, typically with a high threshold to ensure accuracy.

2.5.2. Key Notes:

- Observing conditions play a significant role in plume segmentation. Variations in noise characteristics, scene artifacts, and plume clutter may influence the chosen signal-to-noise threshold.

- Plume segmentation criteria used for emissions quantification should be carefully documented so that others can replicate the work.

- Plume segmentation for visualization should not be used for emissions quantification, as they serve different purposes.

- When dealing with regions with multiple sources or fragmented asset ownership, plume segmentation for attribution may require additional considerations.

Recommendation: It is essential to label plume segmentation products clearly, indicating whether they are intended for visualization or quantification purposes.

2.6. Plume Origin and Asset Attribution
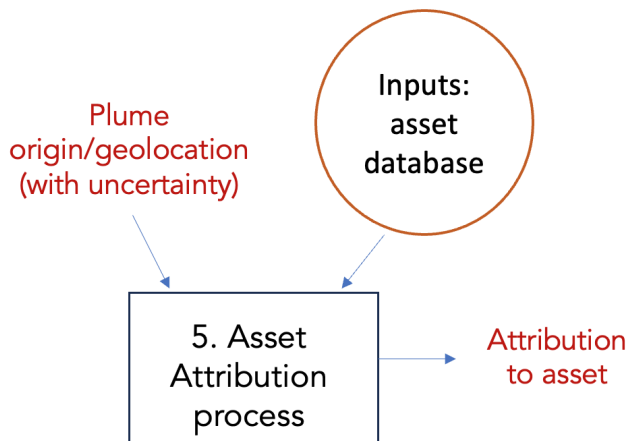
Figure 7. Asset Attribution Step. Concept credit: Dan Cusworth, Carbon Mapper, used with permission.

Once the plume segmentation is complete, the next step is to locate the plume origin. This process is essential for tracing methane emissions back to their source, which is crucial for attribution purposes. Identifying the plume origin typically involves the manual evaluation of various data, such as methane concentration fields, wind direction, and surface imagery.

Currently, plume origin determination uses a manual process across all groups. The types of information considered include:

- Concentration fields or matched filter outputs

- Wind direction (and possibly stability)

- The overall shape of the segmented plume (e.g., preference for cone-shaped plumes)

- Surface imagery, including topographical features and infrastructure data

2.6.1. Key Observations:

- There is a significant variation in common practices for determining the plume origin, such as how it relates to concentration gradients, prior information about emissions in that area, or information on underlying infrastructure.

- Plume origin determination is crucial for attribution work. Practitioners are aware of the sensitivity of this process, with low tolerance for errors.

- Large emission sources with consistent winds are easier to attribute, whereas low wind speeds and smaller emissions complicate origin identification.

- Different practitioners use varying sources of infrastructure information, including high-resolution imagery and infrastructure databases. Publicly available data may be incomplete or outdated, leading some teams to use commercial databases that are still imperfect.

Ideally, plume origin determination is accompanied by an uncertainty estimate. This uncertainty depends on several factors, including the spatial resolution of the measurement system, its pointing accuracy, and georeferencing errors. The uncertainty in the plume origin should be put in context of the ground sampling distance and/or spatial response function in the data products. For example, if the uncertainty is reported in meters, the GSD/SRF for the particular viewing angle and region of the detector of the measurement should also be reported in meters.

A distinction can be made between plume origin and asset attribution. Asset attribution involves identifying a specific piece of equipment as the likely source of the detected emissions. This process typically involves reviewing plume origin data in conjunction with equipment maps and databases. The equipment nearest to the plume origin and most likely responsible for the emissions is identified as the attributed asset. Not all teams include asset attribution in their workflow. They may provide only the geographic location of the plume origin.

Asset attribution can be a source of significant disagreement among practitioners. Much of the disagreement stems from the varying databases teams use for asset identification. Publicly available databases often lack the necessary detail, and different teams may employ different data sources. Furthermore, there is a critical relationship between the GSD of the measurement instrument and the ability to locate assets. In areas where multiple assets are close together, the GSD must be significantly smaller than the spacing between assets for accurate attribution and the wind direction must not be along the line between adjacent assets.

2.7. Detection Rate and Probability of Detection

When assessing the capabilities of methane detection systems, the concepts of detection rate and probability of detection (POD) are essential. This section explores the formal definitions of detection rates and POD, addresses challenges faced in the field, and outlines methods for evaluating these metrics. Additionally, it discusses the issue of false positives and an approach for estimating the emissions detection threshold.

2.7.1. Definition of Detection Rate and Probability of Detection

Detection Rate: In the context of plume detection, detection rate refers to the ratio of true positive detections to number of observations. The detection rate offers some insight into the likelihood that a methane source will be detected by a measurement system. Other ratios of interest are the false positive rate (plume detected when no plume was present), true negative detection rate (no plume detected when none was present), and the false negative rate (no

plume detection when a plume was present). In general, there is very little information available to determine anything other than the true positive detection rate. Controlled release experiments have provided the most insight into detection rates to date.

Probability of Detection (POD): POD is the term used in this community for capturing information about a measurement system's ability to detect methane plumes of various emission rates. The formal definition corresponds to the definition of detection limit provided by the Joint Committee for Guides in Metrology (JCGM VIM, 2012). This definition relies on knowledge of the probabilities of false positives and false negatives. The POD is not a single value, but a set of values as a function of key driving variables.

JCGM Definition of Detection Limit, which corresponds to our term Probability of Detection: The probability of detection is defined as the measured quantity value obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component is β, given a probability α of falsely claiming its presence.

- JCGM Note 1: IUPAC recommends default values of 0.05 for both β and α.

- JCGM Note 2: The term "LOD" (limit of detection) is sometimes used.

- JCGM Note 3: The term "sensitivity" is discouraged when referring to detection limits.

In the application to methane plume detection, term β refers to the false negative rate, and the term α refers to the false positive rate. Note that the determination of the POD curve does not require quantification of plume emissions by the measurement system. If controlled release data is used, the POD curve can be determined from the known emission rates of the controlled release and the detects and non-detects of the measurement system.

2.7.2. Challenges in Implementation

Teams in the methane detection community face various challenges when it comes to evaluating and applying the POD concept:

- Standard practice is to evaluate systems at a 90% POD (β = 0.1), but there is currently no practical way to evaluate α (false positives).

- False positives can vary across systems and interpretation approaches, making standardization difficult. While some practitioners use loose criteria and allow many false positives, others apply stricter quality assurance processes. This inconsistency suggests that future efforts should focus on improving consensus within the community.

- Machine learning approaches may eventually allow for better control of false positives, potentially offering false positive rates and precision-recall metrics.

- Observing conditions—such as scene brightness and clutter—impact both detection performance and false positives. For instance, uncluttered, bright scenes may perform differently from dark, cluttered ones.

- Controlled-release experiments have primarily focused on releasing emissions while the measurement systems are measuring and not creating the no-emissions cases that are needed for testing for false positives and true negative results.

- Note that the term "minimum detection limit" is often misinterpreted as the smallest emissions ever observed. For this reason, the community avoids using it.

To ensure comparability across different systems, teams should document observing conditions alongside their POD evaluations (Figure 8b). Ideally, a standard set of reference conditions (viewing geometry, surface albedo, etc.) would be used across all teams.
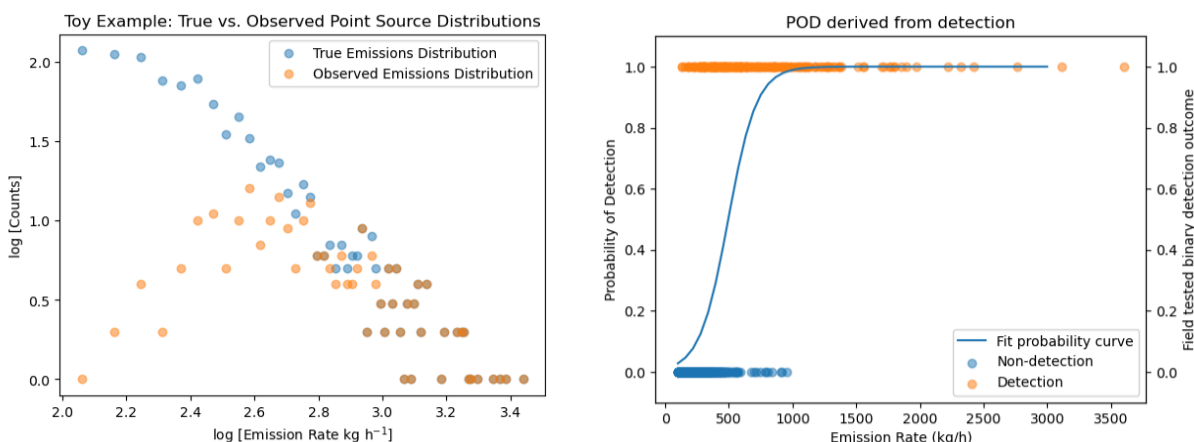
Figure 8. Example of Emissions Distribution Curve (left) and Parametric POD Curve (right). Figure courtesy of Dan Cusworth, Caron Mapper, used with permission.

The left plot of Figre 8 shows the distribution of "true" emissions in a field (blue dots derived from airborne survey described in Cusworth et al., 2025), and the orange dots represent an example that assumes EMIT sampled the basin at the same time and detected plumes following the curve described in Ayasse et al. (2024). A logistic curve is fit using simulated detection vs non-detects, which is shown in the right plot. This is the POD curve. This shows 100% POD after 1000 kg·h-1, but that there are detections as low as near 100 kg·h-1.

Note that detection performance in single-blind testing with a single known source location is not necessarily representative of detection performance in the field (El Abbadi et al. 2024, Kunkel et al. 2023). Expansion of controlled-release experiments to a wider range of observing conditions (surface reflectivity, scene clutter, viewing geometry, etc.) is needed, and experiments conducted in collaboration with industrial facilities.

2.7.3. False Positives

The term "false positive" refers to the detection of plumes that do not exist. There are a number of reasons that false positives might occur, and some of those are illustrated below in Figure 9. For example, clouds or a smoke plume might be incorrectly identified as a methane plume. Surface features with surface reflectance that contrasts the background (such as

calcium carbonate or mineral or solar panels) may be misidentified as a plume. Current practices rely heavily on human interpretation to detect false positives – generally if there is no oil and gas or other potential emitting infrastructure near the plume origin, or the visible or SWIR imagery shows features that explain the apparent presence of a plume, the plume will be discarded. Note that in single-blind controlled methane release testing with a known location, no satellite-based methane sensing system has yet produced a false positive, although this does not preclude the possibility of false positives in the field (Sherwin et al. 2023, 2024a). Machine learning approaches may eventually allow for better control of false positives, potentially offering false positive rates and precision-recall metrics.
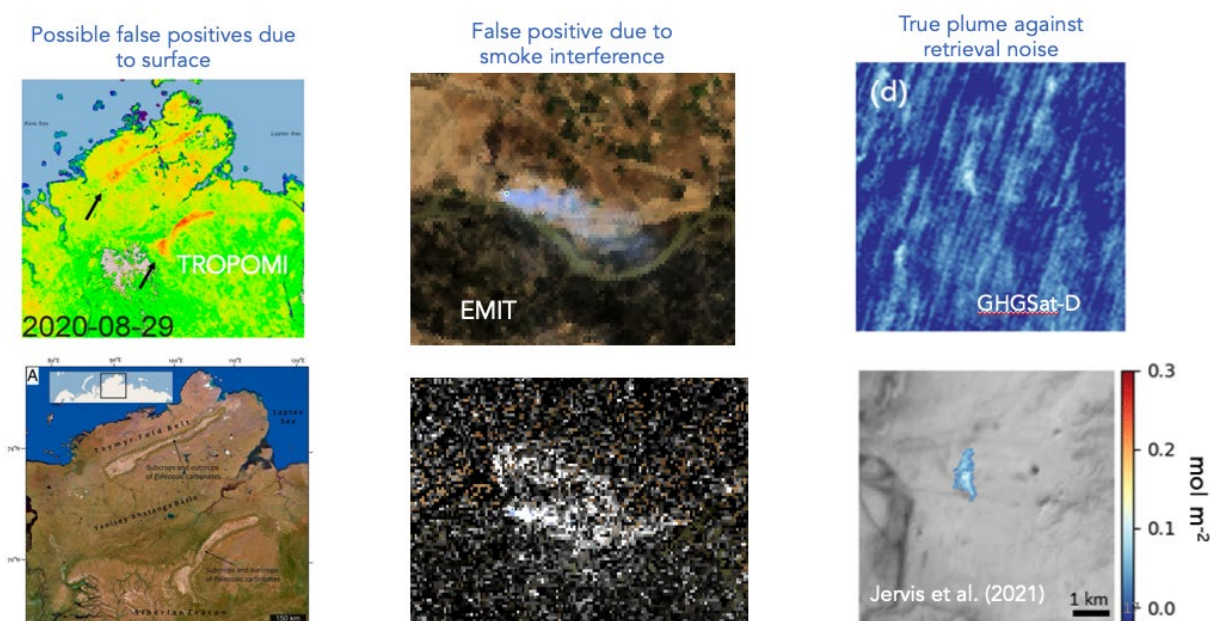
Figure 9. Some Examples of Conditions that Can Lead to False Positive Detections. Credit: Dan Cusworth, Carbon Mapper, used with permission.

2.7.4. Estimated Detection Threshold (in reference to emissions)

In Jacob et al. (2016), the term "detection limit" is used to refer to the lower limit of the measuring interval for emission quantification. To align with the language of the metrology community, we will refer to this as the estimated detection threshold. The measuring interval is defined by the JCGM VIM as "set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions". Prior to having measurement data that can be used to determine the lower limit of the measuring interval and probability of detection curve as per the definitions, an approach to estimate these parameters can be constructed.

Mass balance arguments are used in Jacob et al. (2016) to derive a formulation to determine the estimated detection threshold (lower limit of measuring interval). The Jacob et al. 2016

formula (below) is a simple approximation that can be used to roughly estimate the detection limit of systems and observing condition when there is not enough controlled-release data to empirically determine the POD curve. It is useful to analyze and tabulate estimated detection limit values for many systems and observing conditions, planned and existing.

$$Q_{min} = (M_{CH4} U*W*p*q*\sigma)/(g*M_a) \tag{1}$$

- Q_{min} is the minimum detection limit in Jacob et al., (2016) ($kg \cdot h^{-1}$) [which we will refer to as estimated detection threshold]

- W is the pixel size (m)

- U is the wind speed ($m \cdot s^{-1}$)

- M_{CH4} is the molecular weight methane of methane (0.016 $kg \cdot mol^{-1}$)

- M_a is the molecular weight of air (0.029 $kg \cdot mol^{-1}$),

- p is the dry atmosphere surface pressure (Pa),

- g is the acceleration due to gravity (9.8 $m \cdot s^{-2}$),

- σ is the precision expressed in mol/mol, which is determined from modeled/predicted instrument performance

- q is {2,5} (2 is for used detection, 5 for quantification). These come from the Jacob et al. (2016) definition of detectability as a precision of delta-X/2 and quantification as a precision of delta-X/5, where delta-X is the mean enhancement.

2.8. Emissions Quantification, Common Practices and Open Issues
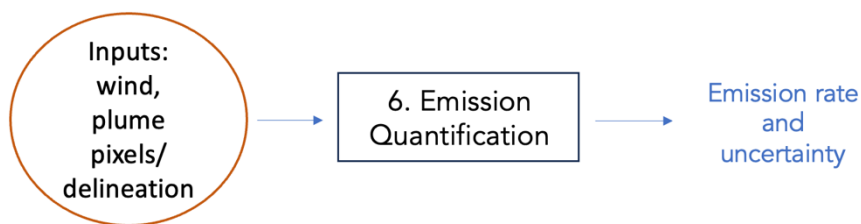
Figure 10. Emission Quantification Process Steps. Concept credit: Dan Cusworth, Carbon Mapper, used with permission.

Emission rate quantification is an important aspect (Figure 10) of methane plume detection and mitigation, because the decision to take mitigation action relies on this information. Emission rate quantification involves estimating the amount of methane being emitted from a source, often through various measurement methods. This section outlines the commonly used techniques, as well as some of the open issues that practitioners encounter when trying to quantify emissions with precision and accuracy.

The two most commonly employed methods for quantifying methane emission rates are integrated mass enhancement (IME) and cross-sectional flux (CSF). These methods have

become the standard for many practitioners due to their effectiveness in different emission scenarios. Other quantification methods are also used in the community (see Figure 11), but they are generally variations or refinements of IME and CSF. These additional methods, along with best practices for their application, will be addressed in later versions of this document as more research is conducted and new techniques are validated.

Gaussian Plume	Local mass balance	Gauss's theorem
$Q = U \Delta \Omega(x,y) \left(\sqrt{2\pi} \sigma_y(x) e^{\frac{y^2}{2\sigma_y(x)^2}} \right)$	$Q = UW\Delta\Omega$	$Q = \oint_s \Omega(s) \vec{U} \cdot \vec{n} \, ds$
Integrated mass enhancement (IME)	Cross-sectional flux (CSF)	Angular width
$Q = U_{\text{eff}} \, \text{IME}/L$	$Q = U \int_a^b \Delta\Omega(x,y) \, dy$	$Q = f(\text{IME}, \theta)$
Computer vision		$Q = \text{NN}(\text{Plume image})$
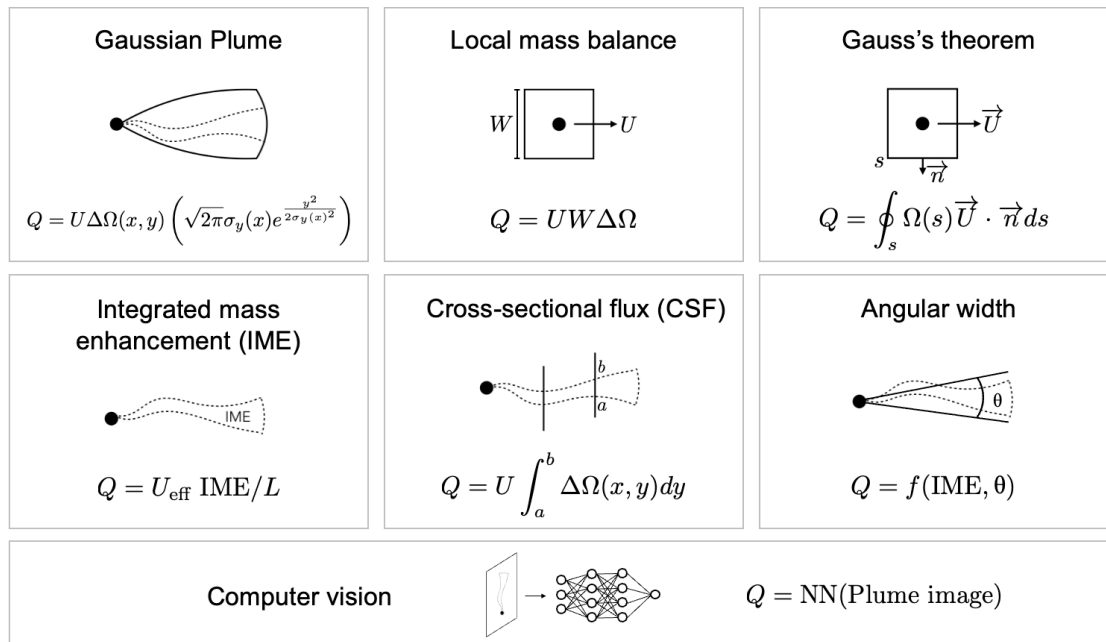
Figure 11. Many Emission Estimate Methods can be Applied. Adapted from Jacob et al. 2022; see reference for details and additional references. Figure courtesy of Daniel Varon, used with permission.

1. Integrated Mass Enhancement (IME): IME is used to calculate methane emissions by integrating the column mass enhancement of methane (kg/m^2) in the detected plume. This method typically involves summing the column mass enhancements across the pixels of the plume and applying information on the plume extent and wind speed. An empirical factor (U_{eff}), discussed more below, is included by some teams. The IME method is less computationally demanding than the cross-sectional flux approach, but it requires accurate input data, especially regarding plume dimensions and wind conditions.

2. Cross-Sectional Flux (CSF): CSF methods estimate the flux of methane by calculating the cross-sectional methane concentration at one or more points along a plume and combining this with wind speed data. This method offers an alternative to IME and is especially useful when the plume edge is well defined and wind data are reliable. The accuracy of CSF can be affected by uncertainties in wind information and assumptions regarding plume structure.

The details of these two approaches, including the equations and term definitions here, taken from Varon 2018 are presented here.

IME equation:

$$Q = \frac{1}{\tau}\text{IME} = \frac{U_{\text{eff}}}{L}\text{IME} = \frac{U_{\text{eff}}}{L}\sum_{j=1}^{N} \Delta\Omega_j A_j.$$

Terms:

- Q – emission rates estimate

- τ – residence time of methane in the detectable plume

- IME – integrated mass enhancement

- U_{eff} – operational parameter related to wind speed

- L – operational parameter that captures plume extent (see more below)

- $\Delta\Omega_j$ – column mass methane enhancement at pixel j

- A_j – area of pixels in plume

Regarding the operational parameters U_{eff} and L: L is a measure of plume extent, which can be interpreted as the length or perimeter of the plume, or the square root of the area of the plume. U_{eff} is an effective wind speed that is derived from a simulation dataset, typically using large eddy simulations (LES). Given a source of wind data (U_{10} for example), where U_{10} is the 10 meter (altitude) wind product, and a choice of definition of L, an empirical relationship between U_{eff} and U_{10} is derived. U_{eff} should not be interpreted as actual effective wind speed, since it is a parameter that can vary based on the definition of L. The simulations are derived for specific instrument configurations (spatial resolution and noise), and also for specific atmospheric and surface conditions, such as value or range of sensible heat flux and mechanical turbulence or the source spatial extent (pure point source versus emissions over a larger area).

CSF equation:

$$Q = \int_{-\infty}^{+\infty} U(x,y)\Delta\Omega(x,y)\mathrm{d}y,$$

From Varon et al. (2018): "By mass balance, the source rate Q must be equal to the product of the wind speed and the column plume transect along the y axis perpendicular to the wind (see CSF eqn). The integral is approximated in the observations as a discrete summation of the product U(x,y) and $\Delta\Omega$(x,y) over the detectable width of the plume." The wind speed in two dimensions is noted as U(x,y). The column enhancement in pixel x,y is referred to as $\Delta\Omega$(x,y).

Varon notes that "A disadvantage of the CSF approach is that the wind U(x,y) is not as well characterized. It must describe some vertical average over the plume extent and there is generally no information on its horizontal variability over the scale of the plume. This may require estimation of an effective wind speed U_{eff} applied to the cross-plume integral C [$kg \cdot m^{-1}$] of the column along the y axis."

2.8.1. Current issues with the IME approach

The application of the IME equation requires L (plume length or plume area), U_{eff}, and IME, or the summed methane enhancement in the pixels included in the plume. L and U_{eff} are effectively co-dependent quantities, where LES calculations are used to develop the relationship between these variables. The relationship between U_{eff} and U_{10} depends on the choice of definitions (L as a plume length or plume area), plume delineation (which pixels are

included), as well as the concentration retrieval (methane enhancement in the included pixels). When used in data analysis, the source of wind information and definition of L should be consistent with the derivation of the U_{10}-U_{eff} relationship through simulations. It is also recommended that the U_{eff} and L definitions used by practitioners should be clearly reported.

A number of studies have employed LES simulations to derive U_{eff} to U_{10} relationships, starting with Varon et al. (2018) for the GHGSat instrument. In later papers, simulations were performed that expanded on the conditions (wider range of sensible heat flux for example) and a range of instrument specifications. U_{eff} to U_{10} relationships for different pixel resolutions, retrieval noise levels, and source types are often similar despite the different calibration setups. There are also linear and log-linear formulations expressing this relationship. In the 1 m·s^{-1} to 5 ms m·s^{-1} range, the linear relationships reported in the literature yield values that may only differ by 5% to 15%, but there is more difference between the linear and log-linear formulations. Major differences in the observing configuration can influence the relationship more strongly. For example, a very different U_{eff} to U_{10} relationship was obtained by Dogniaux et al. (2024) for quantifying emissions from the 2022 Nord Stream pipeline leaks, which could be observed only over the bright area of disturbed surface water and not downwind.

Potentially, a large set of LES simulations covering all observing systems, exploring a range of U_{eff} and U_{10} relationships, could sample the full parameter space, but this is deemed impractical. The options currently considered are to a) perform LES for individual scenes / observing conditions, with ever more complex LES to capture local contributions or b) use one ensemble of idealized LES with a range of conditions and accept a larger source rate uncertainty in exchange for versatility and ease of application (D Varon, private comms). Alternatively, others advocate for developing quantification approaches that do not rely on LES and formulations of U_{eff}. It is an open question if this is achievable with a specific class of measurement systems, perhaps depending on spatial resolution and instrument calibration. We anticipate addressing this topic in more depth in future revisions of this document.

Comparing multiple methods provides a rough cross check. Use of both CSF and IMF quantifications, for instance, accesses the advantages of both methods. CSF is not widely used in operational analysis but should be consistent with IME results if both techniques are well formulated.

To better understand the uncertainties in the calculated Q, the comparison of multiple retrievals should be combined with bottom-up uncertainty quantification of the retrieval equation terms, with practitioners sharing enough information to allow a rigorous assessment of their sensitivities. An emission uncertainty budget should contain terms for all the processing steps that lead to the quantified emission: retrieval, segmentation, wind speed, model used to derive effective wind, etc. We anticipate addressing this topic in more depth in future revisions of this document, including approaches to use controlled release measurements to evaluate uncertainty on Q.

2.9. Data Format and Content Recommendations

Data products from different providers currently have a wide range of formats, units, and terminology, which is a barrier to using them together and intercomparing results. We propose the following framework for data product organization and contents.

1. **Data Content and Metadata**

    a. L1B – Calibrated and geolocated radiance
        i.   The preferred SI unit for radiance data is $W/(m^2 \cdot \mu m \cdot sr)$. Practitioners are also reporting in $\mu W/(cm^2 \cdot nm \cdot sr \cdot nm)$
        ii. The wavelength or wavenumber grid must be included along with the radiance spectrum. Typical units for wavelength are micrometers (μm) or nanometers (nm). Typical units for wavenumber are cm^{-1}.
        iii. Uncertainty on the radiance must also be included. This is not always required for the retrievals discussed here but is needed for data characterization and alternative uses of the measurement data.
        iv. Information about the spectral response functions must be available, although it is not necessarily packaged with each radiance spectrum, as it typically varies with wavelength but not time.
        v.   Similarly, information about the spatial response functions must be available, although it is not necessarily packaged with each radiance spectrum, as it is a characteristic of detector element but not time.
    b. L2B – Whole-scene orthorectified atmospheric retrievals
        i.   This is the output of the concentration retrieval step, such as IMAP-DOAS or a full physics retrieval approach
        ii. The expected data units are $mmol \cdot mol^{-1} \cdot m$ (also written as ppm·m or ($\mu mol \cdot mol^{-1} \cdot m$) (path integrated concentration) or $kg \cdot m^{-2}$ per pixel (path integrated concentration) or column averaged dry air mole fraction (($\mu(\mu mol \cdot mol^{-1}$ or ppm).
        iii. A typical file is a Cloud Optimized GeoTIFFs (COG) - orthorectified (latitude/longitude, projected using WGS 84, EPSG:4326). Any resampling applied should be noted.
    c. L2C – Enhancement maps. These are files that are the same size as the whole scene with the pixels that are considered to be enhanced and identified as separate from the background. If using a matched filter (MF), this may be the first L2 product reported.
    d. L3A and L3B – identified plumes
        i.   These files include geotiffs that are GeoJSON data of the plume outline. The plumes should be specifically labeled as for visualization purposes or for quantification.
        ii. This product includes plume origin location, with uncertainty, attribution if available, and some details of plume length or dimension and uncertainties.

    e. L4A – Emission quantification

        i. This is source emissions in $kg \cdot h^{-1}$, with an uncertainty.

        ii. The wind data (source and value) used and an any relevant conversions (grid interpolations, adjustment for elevation, etc.

        iii. Uncertainty terms and overall uncertainty on emission estimate

2. Data documentation

    a. L1b radiance

        i. Should be geolocated using orthorectified, documented methods

        ii. Radiance calibration procedure (if one exists) should be documented with stated uncertainties and a description of the traceability chain.

        iii. Providers should strive to relate their radiances to a known standards during the operating mission, using on-board calibration systems or measurements of well characterized surface calibration sites such as RadCalNet

        iv. Spectral calibration procedure should be described, including the approach for defining the instrument spectral response function (for multi-channel imagers) or instrument line shape function (for spectrometers), along with the method for calibrating spectral range and dispersion

    b. L2B

        i. Process for transforming radiance to total column abundance or dry air mole fraction or matched filter (MF) results should be documented.

        ii. Strive to connect total column concentration values to standard such as TCCON or COCCON by overflying sites and developing calibration curve.

        iii. Provide precision estimate on total column abundance or dry air mole fraction

        iv. Provide uncertainty estimates on total column abundance or dry air mole fraction

        v. If using MF, all key parameters should be documented in an algorithm theoretical basis document (ATBD) if fixed, per scene if they vary

    c. L2C – detection. This is an area where there is a wide range of practices. Methodology should be described in ATBD or user guide.

    d. L3A and L3B – This is an area where there is a wide range of practices. Methodology should be described in ATBD or user guide.

3. The State of Validation for Point-source Methane Sensing Satellite Systems

3.1. Introduction

Multiple satellite-based systems exist to detect and quantify methane point sources. These methane-sensing systems combine satellite-based observations of multiple column-integrated light spectra (L0 data) with various forms of data analysis to generate geolocated estimates of methane emissions from a site, facility, or location (L4 data).

Because of the focus on detection and quantification of methane emissions, the primary mechanism of validation for such systems has to date been based on observations of single-blind controlled methane releases. We do not discuss traditional methods for validating satellite-based greenhouse gas dry air mole fraction estimates against those from ground-based, airborne, or shipborne in situ sensors or to remote sensing observations from total carbon column observing network (TCCON) or the collaborative column observing network (COCCON) stations. These comparisons are critical for understanding the precision and accuracy of the space-based estimates of column-averaged dry air mole fractions used in regional-scale or global flux inversion models. Although such data are sometimes used as inputs for site-scale methane emissions algorithms, many teams rely on matched filter approaches to identify enhancements, precluding the need for a comprehensive understanding of column-averaged concentration estimates for emissions detection and quantification, which would require a different validation approach.

3.2. Current Controlled-Release Approach for Satellites

The application of controlled-release experiments to validate methane emissions estimates from satellite data has a limited scope with first papers being published based on tests in 2020, 2021, and 2022. Existing tests have implemented a single-blind, known location design with methane emission rates ranging from 0.03 t/·h^{-1} to 7.6 t·h^{-1} (Sherwin et al., 2023; Sherwin et al., 2024a; Darynova et al., 2023). In these tests, an independent testing agent, such as a research institution, conducts metered releases of undisclosed volumes of methane as satellites pass overhead. The satellites collect such measurements over the course of a study period, lasting several weeks to months (or, in the case of Darynova et al. (2023), a single measurement on one day). For each satellite overpass, teams analyzing satellite data then report the presence/absence of emissions (detection) and estimate the amount of methane released (quantification) without access to any operational data from the release.

The test location should be far from potential confounding sources of methane, e.g., oil and gas facilities, large landfills, dairies, as has been the case for major controlled releases conducted so far (Sherwin et al., 2023, Sherwin et al., 2024a). Some satellites can detect plumes well over 1 km from the source (Sherwin et al., 2023). The test location should be instrumented with high-quality wind sensors, especially at 10 m height, because the quantification models used rely on wind speed to estimate flux rate, and therefore, ground truth data on wind speeds should be collected.

The testing agent then compares these detection and quantification reports with metered emission rates. As a best practice, the testing agent then publishes these results (ideally in a peer-reviewed format) in a manner that is independent of the tested technology providers (e.g., without providing the tested parties with some form of veto power over publication of the results). Tests so far have modeled experimental design on the Advancing Development of Emissions Detection protocol for aerial technologies (Bell and Zimmerle, 2020).

Controlled-release testing provides insights into detection capabilities discussed in Section 2, including:

- Providing an upper bound on the smallest emission a given satellite system is capable of detecting

- Determining the presence/absence of false positive detections (reports of emissions when none were present)

- Characterizing the range of emission sizes a technology system can detect with a given level of reliability

Note that no false positives have been observed in any single-blind test of a satellite-based methane sensing system conducted at the time of this writing (Sherwin et al., 2023; Sherwin et al., 2024a; Darynova et al., 2023), although this does not rule out the possibility of false positives in the field. In addition, existing controlled-release tests have only a single point source, while there are many examples in the field of multiple sources coalescing into a single plume (Sherwin et al., 2023; Sherwin et al., 2024a; Darynova et al., 2023, Thorpe et al., 2023)

The characterization of the emissions sizes a system can detect requires a comparatively large sample size. This is because in order to determine the reliability of detection at a given mass flow rate, multiple releases of that rate must be conducted so that performance can be assessed (e.g., 17% of emissions of rate 100 kg/h to 150 kg/h were detected). This level of test coverage has not been achieved in tests to date. Across all single-blind tests conducted so far, no satellite-based methane sensing system has more than 15 valid measurements (Sherwin et al., 2023; Sherwin et al., 2024a; Darynova et al., 2023).

3.3. Current Controlled-Release Approach for Aircraft Systems

Similar tests aiming to provide detailed characterization of lower detection capabilities of airplane-based methane remote sensing systems typically require on the order of 100 data points or more (El Abbadi et al., 2024; Bell et al., 2022; Sherwin et al., 2019). As a result, although existing studies provide some insight into the lower detection capabilities of the tested satellite-based methane sensing systems, additional testing is needed to provide statistically robust characterization of the detection probability curve.

These tests also provide insight into quantification capabilities, including:

- Characterizing any bias in quantification volumes across measurements

- Characterizing uncertainty associated with a given measurement

3.4. Findings to Date

In the airplane-based methane remote sensing literature, characterizing a quantification error distribution is possible with a sufficient number of measurements (El Abbadi et al., 2024). The more measurements collected, the greater insight one can gain into the tested system's uncertainty. At present, the largest number of nonzero measurements for a given team analyzing a single satellite is 6 across multiple tests, not enough for a detailed characterization of quantification uncertainty (Sherwin et al., 2023; Sherwin et al., 2024a; Darynova et al., 2023).

Due to these sample size limitations, tests so far have focused on characterizing the quantification capabilities of a suite of satellite-based methane sensing systems across multiple satellites and analysis teams. These results provide a rough assessment of the maturity of the field of satellite-based point source quantification, rather than assessing the quantification bias or uncertainty of an individual satellite-based methane sensing system. Results so far suggest that satellite-based point source quantification approaches tend to be roughly unbiased, with individual measurements subject to a level of uncertainty that is qualitatively similar to that observed in many aircraft-based methane remote sensing systems, with 55% to 75% of measurements falling within ±50% of the metered value (Sherwin et al., 2023; Sherwin et al., 2024a; El Abbadi et al., 2024; Bell et al., 2022).

In most methane remote sensing algorithms, the estimated emission rate is modeled as linearly related to estimated wind speed, such that an overestimate of 2x in wind speed will increase the estimated emission rate by 2x. Because on-the-ground empirical wind speed measurements are generally not available in satellite-based methane remote sensing, it is common practice to rely on wind reanalysis data.

Two tests so far have conducted a second stage of blinded testing to evaluate the effect of wind speed assumptions on quantification performance (Sherwin et al., 2023; Sherwin et al., 2024a). After teams have submitted fully blinded detection and quantification estimates, they are then provided with ground-based wind speed measurements (typically from an on-site 10 m ultrasonic anemometer). Teams then have an opportunity to submit updated emission rate estimates incorporating the empirical measured wind data. These wind-unblinded estimates demonstrate the significant uncertainty introduced into satellite-based methane quantification estimates, with the R^2 from a fixed-intercept ordinary least squares regression rising from 0.585 to 0.772 (Sherwin et al., 2024a), suggesting a much-improved linear fit to the combined data from all tested satellites and teams when direct wind measurements are available.

In addition to characterizing detection and quantification performance, single-blind controlled methane release tests also provide insight into the sensitivity of different approaches to environmental factors such as clouds. No satellite system of which we are aware can currently detect methane through fully overcast conditions. However, recent single-blind tests revealed substantial variability in detection performance across satellites and teams under partially cloudy conditions (Sherwin et al., 2024a).

3.4.1. Limitations of Known-Location Single-Blind Testing

Ideally, validation efforts would provide a clear picture of a satellite-based methane sensing system's detection and quantification capabilities over a wide range of landscapes, environmental, and meteorological conditions.

Single-blind tests conducted so far are an important first step in this direction. However, they have several important limitations that should be addressed in future campaigns:

- Single location: Tests conducted so far have been performed at a single location, generally at a location with favorable conditions for methane detection (e.g., a desert environment with few nearby structures, low cloud cover, and relatively simple scene complexity).

- Small sample size: Because a given methane-sensing satellite will only pass overhead every 1 to 16 days, past campaigns of 3 to 8 weeks are not able to collect sufficient data points to rigorously characterize the detection and quantification capabilities of individual systems (Sherwin et al., 2023). This infrequent revisit time makes it costly to collect large sample sizes.

- Known location: In tests conducted so far, participating teams are aware of the test location and the testing period. As a result, analysis teams may be able to identify smaller methane emissions based on data that might not pass quality control if captured under other circumstances.

- Single release point: In the field, what appears to a satellite as a single plume may in fact be a complex of nearby sources, which could be more challenging to accurately quantify.

One approach that can partially mitigate the above issue is to require full-field retrieval images as well as masked plume images for all measurements, including detections, non-detections, measurements excluded due to data quality issues, and measurements excluded from analysis due to prior disclosure of emissions schedules (e.g., if a team was notified that there would not be releases on weekends but collected measurements on weekends anyway). Full-field retrieval images give additional insight into whether an identified plume is clearly distinguishable from the background. Asking for cloud and artifact maps (e.g., due to water bodies) can also assist in the interpretation of full-field retrieval images.

3.4.2. Alternative Validation Methods

In addition to blinded controlled methane release testing, multiple other approaches exist for validation of facility-scale methane remote sensing. This is particularly true for validating detection capabilities. Prominent methods include simulation experiments, multi-scale measurement intercomparison, and operator estimation of detected emissions.

Simulation experiments attempt to assess the performance of satellite-based methane sensing, particularly detection capabilities, adding a simulated methane signal to measured or simulated data. One such approach uses analytic estimation of minimum detection capabilities as a

function of sensor and mass balance characteristics (Jacob et al. 2016). Several recent studies use computational models to add simulated high-fidelity methane plumes to actual spectral data collected via satellite, collected in locations that appear to have no detectable methane sources (Cusworth et al. 2019, Guanter et al. 2021, Gorrono et al. 2023). These relatively low-cost approaches provide valuable insight into the theoretical capabilities of satellite-based methane sensing systems, including those that are not yet operational. These plume simulation approaches enable assessment of the plume detection ability of a given sensor and processing chain under different acquisition conditions and flux rate ranges. They also facilitate the intercomparison of retrievals from different missions by embedding the same plume on different datasets. However, the accuracy of the simulations must be assessed (e.g., by comparing simulated plumes with real ones from controlled releases) in order to ensure the reliability of these methods.

Another validation method for methane remote sensing systems, including satellites, is multi-scale measurement intercomparison. In synchronous instrument cross comparison, two or more methane sensing systems measure the same source at roughly the same time. Duren et al. (2019) employ this approach, comparing simultaneous measurements from the airborne visible infrared imaging spectrometer - next generation (AVIRIS-NG) imaging spectrometer and the Scientific Aviation aerial mass balance technique to estimate emission rates both from controlled releases and unmetered industrial sources. This method, which generally focuses on validating quantification capabilities, can have lower cost than single-blind controlled releases, but with substantially greater uncertainty given that most methane-sensing technologies have quantification error with 95% confidence intervals exceeding +-50% (El Abbadi et al. 2024).

A separate asynchronous instrument intercomparison approach, focused on characterizing detection capabilities, relies on comparing field statistics of emissions detected by two instruments with unbiased quantification in a similar region over a similar time period, e.g., within an oil and gas basin over one year. One can then estimate the detection capabilities of the less sensitive instrument by comparing the frequency with which emissions of a given size were detected by each instrument. Kunkel et al. (2023) employ such an approach to estimate the detection capabilities of the AVIRIS-NG and Global Airborne Observatory instruments in the Permian basin. These relatively low-cost field statistics-based approaches have the potential to overcome some of the limitations of known-location single-blind controlled-release testing, as it removes the possibility of artificially increasing detection sensitivity during the test by leveraging foreknowledge of the site location and the existence of an ongoing test. However, such an approach requires both a sufficiently large dataset collected by multiple instruments in a consistent region, as well as careful statistical analysis to ensure that the datasets from each instrument are indeed intercomparable (e.g., that the group of assets measured by each instrument have similar characteristics).

Operator estimation of detected emissions can assist in validation of the quantification capabilities of satellite and airborne remote sensing systems. This approach requires low-latency emissions data sharing with an asset operator that can either dispatch or host a post-detection survey team, as in Zimmerle et al. (2024), or analyze operational and maintenance logs to identify the event and bound its magnitude. Due to the intermittent character of many

methane sources, the true emission rate found via post-detection survey may vary substantially from the rate at the time of initial detection (and may have fallen to zero). Analysis of operational and maintenance logs can provide valuable bounds on the emission rate of a detected source, as demonstrated for the Nord Stream II pipeline explosion in **Poursanidis et al. (2024)**.

None of the approaches outlined in this report can, by itself, answer all major questions about the capabilities of methane-sensing satellite systems. Each of these methods can provide valuable and complementary insight toward our collective understanding.

3.4.3. Priorities for Future Testing

- Longer test duration to increase sample size and capture seasonality

- Multiple test locations in varied landscapes and environmental conditions

- Offshore or marshland environments

- Unknown location testing, modeled on the experimental design described in Johnson et al. (2021)

4. Overview: Quality Assessment for Reporting Column Amounts or Column Enhancements

In recent years, the increasing range of applications of earth observation (EO) data products and availability of low-cost satellites has resulted in a growing number of commercial EO satellite systems, developed with a view to deliver end-to-end information services, many of which sense the atmospheric domain. This evolution in the marketplace has led to increasing interest from space agencies in the acquisition of commercial EO data products, as they may provide complementary capabilities and services to those they currently offer.

To ensure that decisions on commercial data acquisitions can be made fairly and with confidence, there is a need for an objective framework with which their data quality may be assessed. The ESA Earthnet Data Assessment Pilot (EDAP) project therefore defines this EO mission quality assessment framework for commercial satellite missions in the optical, SAR, and atmospheric domains. Presented here is the latest evolution of this framework for atmospheric missions that provide measurements of greenhouse gas (GHG) atmospheric columns at facility scale (≈10 meters to 100 meters) and corresponding estimates of emissions from these column amounts.

For this document we use the nomenclature "column amount" to describe the atmospheric measurement of interest. However, "column enhancements" are also reported by this class of instruments where the enhancement is relative to nearby methane column values that are found to represent "background" levels for the region of interest. Additional product files containing uncertainties, albedo, quality flags, and enhanced concentrations that are determined to be part of a methane plume may also be reported. The subsequent section (Section 5) focuses on the methane emissions estimates made available from these measurements, in particular using the file containing plume enhancement values.

4.1. EO Mission Quality Assessment Framework Summary

This section outlines the EO mission quality assessment for atmospheric column data products. The evaluation is primarily aimed at verifying that mission data has achieved the claimed mission performance and, where applicable, reviews the extent to which the missions follow community best practice in a manner that is "fit for purpose".

The approach taken to assess data product quality is based on the QA4EO principle (QA4EO Task Team 2010) and builds on the structure and reporting style developed in other similar work (e.g., Nightingale et al., 2019). This quality assessment framework, developed within the ESA Earthnet Data Assessment Pilot (EDAP) project, aims to build on the experience of this previous work targeting the satellite Cal/Val context. The assessment itself is conducted in two parts, as follows:

- Documentation Review – review of mission quality as evidenced by its documentation

- Detailed Validation – quantitative assessment of product compliance with stated performance

Figure 12. Supply Chain Summary for Level 1 (calibrated radiances), Level 2 (atmospheric column) and Level 4 (emission).

These parts of the assessment, along with their grading criteria, are described in Sections 3 and 4, respectively. The activities are divided into sections and subsections constituting each of the different aspects of data product quality that are assessed and graded. Assessment results are provided in a separate Quality Assessment (QA) Report and are also summarized in a color-coded Product Evaluation Matrix.

It is expected that all relevant mission information needed to perform the assessment would be available to all users, however it is understood that confidentiality may be required for some aspects of a mission. Where this is the case, it will be indicated as confidential in the quality assessment report. In general, pertinent key conclusions of confidential documentation should nevertheless be published openly.

4.1.1. Mission Data Supply Chain Assessment Overview

The specific atmospheric column data product assessment outlined in this document forms part of a wider supply chain assessment summary (Figure 12). This overview matrix encompasses documentation review and detailed validation assessments for all data processing steps for a given atmospheric mission, including calibrated radiances (Level 1B), retrieved atmospheric column products (Level 2), and further derived emissions (Level 4), if applicable.

For each of the rows of the Supply Chain Summary in Figure 12, a full set of data product quality assessment guidelines exist in the same format as shown in this document for the atmospheric column products.

To ensure a complete and transparent quality assessment, EO missions yielding an atmospheric column or enhancement data product must also include some form of documentation and detailed validation assessment for the associated L1B calibrated radiance product used to retrieve the L2 atmospheric product. Although it is appreciated that this may not be possible to a full extent in every case, this section simply recommends assessment at all product processing

levels where possible. Ideally, the L2 atmospheric product is reproducible from the reported L1B data and associated documentation (e.g. ATBDS and product user guides).

4.1.2. Quality Assessment Report

The quality assessment (QA) for a given atmospheric column product is reported using the QA Report template. The template ensures consistency of reporting and facilitates comparison between the assessments of similar missions. The QA report covers each section of analysis, providing more detailed information, and a completed mission product evaluation matrix (see following subsection) presenting the results of each sub-section of analysis in a color-coded table.

4.1.3. Product Evaluation Matrix

The product evaluation matrix provides a high-level color-coded summary of the quality assessment results. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. Subsection grades are indicated by the color of the respective grid cell, which are defined in the key. A padlock symbol in the corner of a given cell indicates that the information used to assess the respective subsection is not available to the public. The reporting of assessment results is divided between two evaluation matrices, as follows:

- Summary Product Evaluation Matrix
- Detailed Validation Maturity Matrix

These matrices are described below.

4.1.3.1. Summary Product Evaluation Matrix

The Summary Product Evaluation Matrix is shown in Figure 13. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. The matrix on the left (in dark blue) summarises the results of the Documentation Review, while the additional column on the right (in light blue) summarises the results of the Detailed Validation. The Validation Summary column is separated from the main table to make clear the results can come from multiple assessment sources.

Data Provider Documentation Review		
Product Information	**Metrology**	**Product Generation**
Product Details	Metrological Traceability Documentation	Atmospheric Column Retrieval Algorithm
Availability & Accessibility	Uncertainty Characterization	Geometric Processing
Product Format, Flags & Metadata	Ancillary Data	Mission Specific Processing
User Documentation		

Validation Summary
Atmospheric Column Validation Methodology
Atmospheric Column Validation Results
Geometric Validation Method
Geometric Validation Results

Key
Not Assessed
Not Assessable
Basic
Good
Excellent
Ideal

Not Public

Figure 13. Summary Product Evaluation Matrix.

4.1.3.2. Detailed Validation Maturity Matrix

The Detailed Validation Maturity Matrix (Figure 14) provides more complete reporting of analysis contributing to the Validation Summary – breaking down the validation methodologies used and the results. This section is aimed at the more technically focused reader. Since, for a given mission, multiple validation studies may be performed – for example, by the mission/vendor and/or by independent assessors – there can be multiple Detailed Validation Maturity Matrices produced and reported. Detailed evaluation (right side) should be performed first and the grades used generate the validation summary (left side).

Atmospheric Column				
Validation Summary		**Detailed Validation**		
Atmospheric Column Validation Methodology	←	Validation Dataset	Validation Method	Validation Completeness
Atmospheric Column	←	Validation Results Compliance		

Key
Not Assessed
Not Assessable
Basic
Good
Excellent
Ideal
🔒Not Public

Validation Results				
Geometric Validation	←	Sensor Spatial Response Method	Absolute Positional Accuracy Method	Temporal Stability Method
Geometric Validation Results	←	Sensor Spatial Response Compliance	Absolute Positional Accuracy Compliance	Temporal Stability Compliance

Figure 14. Validation Maturity Matrix, showing the Validation Summary column from the Product Evaluation Matrix.

4.2. Approach to Grading

The assessment framework is aimed at verifying the claimed mission performance, and to assure that the mission follows community best practice to an extent that is "fit for purpose". The grading criteria for each category are determined based on a logical interpretation of this principle. For example, pre-launch calibration quality grading is based on the comprehensiveness of activity with respect to the target instrument performance.

Grades of Basic, Good, Excellent, or Ideal may be given. The Ideal grade level is generally reserved to provide recognition for achieving the highest standard of quality with respect to community best practice. This high bar of quality may be aspirational but is the benchmark that EO data providers should aim for. Note that a grade of Basic can be considered acceptable in a given context. Grading criteria for each box of the matrix are described in Sections 4.3 and 4.4.

Additionally, a subsection may also indicate Not Assessable or Not Assessed. These cover the cases where certain aspects of product quality will not be assessed – either because there is insufficient information available to make an assessment, or because it is out of scope of the assessment.

4.3. Data Provider Documentation Review

In this section we provide detailed guidelines for Data Provider Documentation Review. This assessment aims to review mission quality as evidenced by its documentation. It is divided into the follow sections:

- Product Information
- Metrology
- Product Generation

In the following we look at each of these sections in turn and discuss the grading criteria.

The results of the Documentation Review are reported on the left portion of the Summary Product Evaluation Matrix (Figure 13). This portion is shown in Figure 15.

Data Provider Documentation Review		
Product Information	Metrology	Product Generation
Product Details	Metrological Traceability Documentation	Atmospheric Column Retrieval Algorithm
Availability & Accessibility	Uncertainty Characterization	Geometric Processing
Product Format, Flags & Metadata	Ancillary Data	Mission Specific Processing
User Documentation		

Figure 15. Data Provider Documentation Review Matrix.

4.4. Product Information

The Product Information section covers the top-level product descriptive information, product format, and the supporting documentation. Its subsections are now defined.

4.4.1. Product Details

Certain basic descriptive information (metadata) should be provided with any EO data product and is required for assessment of all mission domains. This required information is as follows:

- Product name

- Sensor name

- Sensor type: Describe sensor design type, e.g., multi-channel, hyperspectral, interferometer etc., and spectral domains, e.g. visible (VIS), near infrared (NIR), shortwave infrared (SWIR), thermal infrared (TIR).

- Mission type: Either single satellite or constellation of a given number of satellites.

- Mission orbit: For example, Sun Synchronous Orbit with Local Solar Time.

- Product version number

- Product ID

- Processing level of product

- Spatial coverage

- Point of contact (Responsible organisation, including email address)

- Product access (e.g., URL, DOI if applicable)

- Restrictions for access and use, if any

Table 1 shows how provision of data product information relates to its grade for this subsection of the quality assessment.

Table 1. Product Information > Product Details – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside of the scope of study.
Not Assessable	Relevant information not made available.
Basic	Many pieces of important information are missing.
Good	Some pieces of important information are missing.
Excellent	Almost all required information available.
Ideal	All required information available.

4.4.2. Availability & Accessibility

This is about how readily the data are available to those who wish to use them. It does not necessarily require cost-free access but is more about following the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles for scientific data management and stewardship (Wilkinson et al. 2016), which provide valuable principles for all data applications. These state that:

- Data should be findable:
    - Metadata and data are assigned a globally unique and persistent identifier
    - Data are described with rich metadata
    - Metadata clearly and explicitly include the identifier of the data it describes
    - Metadata and data are registered or indexed in a searchable resource
- Data should be accessible:
    - Metadata and data are retrievable by their identifier using a standardized communications protocol
    - The protocol is open, free and universally implementable
    - The protocol allows for an authentication and authorization procedure where necessary
- Data should be interoperable:
    - Metadata and data use a formal, accessible, shared and broadly applicable language for knowledge representation
    - Metadata and data use vocabularies that themselves follow FAIR principles
    - Metadata and data include qualified references to other (meta)data
- Data should be reusable:

o   Metadata and data are richly described with a plurality of accurate and relevant attributes

o   Metadata and data are released with a clear and accessible data usage license

o   Metadata and data are associated with detailed provenance

o   Metadata and data meet domain-relevant community standards

Table 2 shows how a data product's provision of the above information relates to the grade it achieves for this sub-section of the quality assessment.

Table 2. Product Information > Availability and Accessibility – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study
Not Assessable	Relevant information not made available
Basic	The data set does not appear to be following the FAIR principles
Good	The data set meets many of the FAIR principles and/or there is an associated data management plan that shows progress towards the FAIR principles
Excellent	The data set meets many of the FAIR principles and has an associated data management plan and is available either free of cost or through an easy-to-access commercial license
Ideal	The data set fully meets the FAIR principles and has an associated data management plan and is available either free of cost or through an easy-to-access commercial license

4.4.3. Product Format, Flags, and Metadata

An important aspect of EO data products that ensures ease of access to the widest variety of users is their format. Product metadata and flags offer users important extra layers of useful descriptive information, in addition to the measurements themselves, that can be crucial to their analysis.

In the ideal case, the product format would meet the appropriate Committee on Earth Observation Satellites (CEOS)-Analysis Ready Data (ARD) metadata guidelines (CEOS-ARD 2021) requirements. In the case where such a standard does not exist, product format is graded based on the following:

- the extent to which it is documented

- whether a standard file format is used (e.g., NetCDF)

- whether it complies with standard variable, flag, and metadata naming conventions, such as the Climate and Forecast (CF) metadata conventions (Eaton et al. 2020), or, for data from the European Union, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive (European Parliament and Council of the European Union 2007)

- whether flags and metadata provide an appropriate breadth of information

If a product is derived from a constellation of satellites, the specific satellite used should be included in the product metadata.

Table 3 shows how a given EO data product should be graded for its format.

Table 3. Product Information > Product Format, Flags and Metadata – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Non-standard, undocumented data format.
Basic	Non-standard or proprietary data format, or poorly documented standard file format. Minimal useful metadata or data flags provided.
Good	Data exist in a documented standard file format. Non-standard naming conventions used. Includes a good set of documented metadata and data flags.
Excellent	Data are organized in a well-documented standard file format, meeting community naming convention standards. Comprehensive set of metadata and data flags.
Ideal	Analysis Ready Data standard if applicable, else as Excellent.

4.4.4. User Documentation

Data products should be accompanied with the following minimum set of documentation for users, which should be regularly updated as required:

- Product User Guide/Manual (PUG/PUM)

- Algorithm Theoretical Basis Document (ATBD)

It may be for a given mission that in place of these documents some combination of articles, publications, webpages, and presentations provide a similar set of information. For the highest grades however, they should be presented as a formal document, since users should not be expected to search for the information. The QA4ECV project provides guidance for the expected contents of these documents (INSPIRE Drafting Team Metadata and European Commission Joint Research Centre 2013), (INSPIRE Thematic Working Group Orthoimagery 2013), which they can be evaluated against.

Table 4 describes how the assessment framework grades a products user documentation.

Table 4. Product Information > User Documentation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No user documentation provided or documentation out-of-date.
Basic	Limited PUG available, no ATBD. Information is up-to-date.
Good	Some PUG and ATBD-type information available. These may be formal documents or from multiple sources. Documentation is up-to-date.
Excellent	PUG meets QA4ECV standard, reasonable ATBD. Documents are up-to-date.
Ideal	PUG and ATBD available meeting QA4ECV standard. Documents are up-to-date.

4.5. Metrology

Metrology is the science of measurement. This section covers the aspects of the mission related to measurement quality, including calibration, traceability, and uncertainty. The Metrology subsections are now defined.

4.5.1. Metrological Traceability Documentation

Traceability is defined in the vocabulary of metrology (VIM) (JGCM 2012) as a,

"property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty"

and reinforced in the QA4EO procedures. Traceability is therefore a key aspect of achieving reliable, defensible measurements. In this definition an important part of measurement traceability is highlighted – that it is well documented. This of course must be the case for EO data products too.

Various diagrammatic approaches have been developed to present the traceability chains for EO data products (e.g., the QA4ECV guidance, which includes a traceability chain drawing tool (Scanlon 2017)). Such a diagram should be included in the documentation for every EO mission. The FIDUCEO project has provided guidance for a more detailed measurement function centered "uncertainty tree diagram" which is ultimately more suitable for most examples of EO data processing and should be the aspiration for missions in the future (Datla et al., 2011).

Table 5 shows how the assessment framework grades the metrological traceability documentation, based on its completeness.

Table 5. Metrology > Metrological Traceability Documentation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No traceability chain documented.
Basic	Traceability chain diagram and/or uncertainty tree diagram included, missing some important steps.
Good	Traceability chain and/or uncertainty tree diagram documented identifying most important steps and sources of uncertainty.
Excellent	Rigorous uncertainty tree diagram, with a traceability chain documented, identifying all reasonable steps and accompanying sources of uncertainty.
Ideal	Rigorous uncertainty tree diagram and traceability chain documented, identifying all reasonable steps and accompanying sources of uncertainty. Establishes traceability to SI.

4.5.2. Uncertainty Characterization

To ensure measurements are both meaningful and defensible, it is crucial that they include rigorously evaluated uncertainty estimates. A comprehensive description of how to evaluate

sources of uncertainty in a measurement and propagate them to a total uncertainty of the final measurand is provided by the metrological community in the Guide to the Expression of Uncertainty in Measurement (GUM) (JCGM 2008).

The application of Earth Observation metrology has progressed greatly in recent years. Increasingly, providers of operational and reprocessed data products are applying different approaches to evaluate and distribute metrologically rigorous error-covariance at the per-pixel level, as required by climate studies. For example, ESA's Sentinel-2 mission has developed an on-the-fly, pixel-level uncertainty evaluation tool (Gorroño et al., 2017). There have also been some initiatives, like the previously mentioned FIDUCEO project, that have applied metrology principles to historical sensor data records (Mittaz, Merchant, and Woolliams 2019).

With that said, it is typical for uncertainties (or performance estimates) to be evaluated in a manner that does not comply with the GUM. For example, uncertainties in optimal estimation retrieval algorithms are propagated within the retrieval itself (within prior and measurement error covariance matrices), so "traditional" GUM approaches to uncertainty propagation are not strictly applicable here. Furthermore, many trace gas column product uncertainties are simply derived primarily as the spread and offset of observations relative to validation data (e.g., the Total Carbon Column Observation network [TCCON]). We therefore do not specify a strict requirement for GUM approaches in product uncertainty analysis for higher assessment grades, but it is required that the explanation of the approach that was used is clear.

Table 6 shows the uncertainty characterization grading under the assessment framework.

Table 6. Metrology > Uncertainty Characterisation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No uncertainty information provided.
Basic	Uncertainty established by limited comparison to measurements by other sensor/s.
Good	Limited use of rigorous uncertainty estimation approaches, and/or, an expanded comparison to measurements by other sensors. Most important sources of uncertainty are included.
Excellent	Metrologically rigorous approach used to estimate measurement uncertainty, all important sources of uncertainty are included. Uncertainty per pixel provided.
Ideal	Metrologically rigorous approach used to estimate measurement uncertainty, including a treatment of error-covariance. Per pixel uncertainties in components, e.g., random systematic – as appropriate for the error-correlation structure of the data.

4.5.3. Ancillary Data

Throughout the processing chain there may be a requirement for external input data, for example, a priori atmospheric state information, or reference data for algorithm tuning. The ancillary datasets used during the processing should be identified to the user (where possible due to commercial sensitivity). Ideally this should be traceable on a per product level.

Ancillary datasets must be of a sufficient quality, including the application of suitably rigorous metrology, for example, in the form of SI traceability.

The suitability of the ancillary data for its application must also be considered, with respect to the mission's stated performance requirements. For example, the quality, size and representativeness of algorithm input data. The requirements will be specific to the retrieval method used and may require some expert judgement.

Table 7 shows how the ancillary data are graded under the assessment framework.

Table 7. Metrology > Metrology > Ancillary Data – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Use of ancillary data undocumented.
Basic	Ancillary data used in product generation, specified to some extent, though incomplete. Not entirely of a sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.
Good	Ancillary data used in product generation, specified, though not necessarily on a per product basis. Mostly of a sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.
Excellent	Ancillary data used in product generation, fully specified per product, and traceable. Ancillary data used are of sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.
Ideal	Ancillary data used in product generation, meets the Excellent criteria, and are traceable to SI where appropriate.

4.6. Product Generation

The Product Generation section covers the processing steps undertaken to produce the data product. This primarily concerns the retrieval algorithm used to derive atmospheric column quantities from satellite instrument measurements, and further processing that may be required post-retrieval.

4.6.1. Atmospheric Column Retrieval Algorithm

There are typically a variety of potential retrieval methods available to derive atmospheric column products, such as optimal estimation-based inverse methods, proxy retrieval methods, or band differencing methods applied to hyperspectral/multispectral instruments (e.g., Sentinel-2, (Gorroño, et al., 2023)). The retrieval methods vary in model complexity and computational efficiency – resulting in higher- or lower-quality final products.

The L2 atmospheric column retrieval method should be of a sufficient quality that is "fit for purpose" within the context of the mission's stated performance across all stated use cases (e.g., scene types). What the retrieval method requires is specific to a given variable's retrieval methods and will require a degree of expert judgement.

Table 8 shows how the assessment framework grades the retrieval algorithm used to generate L2 products.

Table 8. Product Generation > Atmospheric Column Retrieval Algorithm – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Retrieval algorithm not documented.
Basic	Retrieval algorithm somewhat documented. Retrieval algorithm too simple to be judged "fit for purpose" in terms of the mission's stated performance.
Good	Retrieval algorithm documented. Retrieval algorithm judged "fit for purpose" in terms of the mission's stated performance. The documentation includes the algorithm for generating the column enhancement and plume mask.
Excellent	Retrieval algorithm well documented. Retrieval algorithm is "fit for purpose" in terms of the mission's stated performance. The documentation includes the algorithm for generating the plume mask. The algorithms are published and peer reviewed.
Ideal	In addition to meeting the excellent criteria, the full uncertainty budget for the column retrieval algorithm and plume mask generation are described.

4.6.2. Geometric Processing

Several different geometric processing methodologies may be applied to optical imagery data depending on the application of the data product. These may include selection of the Earth model (National Imagery and Mapping Agency, 2000), terrain surface model (Wolfe et al., 2013), correction to ground control points (GCPs), resampling or orthorectification amongst others. Processing may vary between products for a given mission, for example, based on number of available GCPs or geolocation references (Gutman et al., 2013; Storey, Choate and Lee, 2014; Dechoz et al., 2015).

The geometric processing should be of a sufficient quality that is "fit for purpose" within the context of the mission's stated performance for all mission products. Again, this constitutes a technical review of the ATBD from the data provider.

Table 9 shows how geometric processing is graded.

Table 9. Product Generation > Geometric Processing – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Geometric processing not fully documented.
Basic	Geometric processing documented. Missing all or part of the calibration parameters. Calibration algorithm too simple to be judged "fit for purpose" in terms of the mission's stated performance. Confidence in the calibration quality is minimal.
Good	Geometric processing documented. Missing part of the input calibration parameters. Reasonable retrieval algorithm used. Confidence in the calibration quality is considered sufficient.
Excellent	Geometric processing documented. All input calibration parameters exist. Methodology used is considered "fit for purpose" in terms of the mission's stated performance for all expected use cases. Quality flags indicate good geometric accuracy with less than 5% exceptional.
Ideal	Geometric processing well-documented. State-of-the-art methodology used, easily "fit for purpose" in terms of the mission's stated performance. Quality flags indicate excellent geometric accuracy.

4.6.3. Mission-Specific Processing

Additional processing steps are separate from the main retrieval processing. These may include processes like the generation of quality or cloud masks. Additional processing steps must themselves be assessed for quality based on their "fitness for purpose" in the context of the mission.

In the case of additional processes where the measurement data themselves are transformed in some manner, such as orthorectification, the uncertainties from the measurement data must be propagated, as well as introducing appropriate additional uncertainty components caused by the processing itself. This is required for the uncertainties to remain meaningful.

Each additional processing step should be separately assessed based on the criteria described in Table 10, and then a combined score determined.

Table 10. Product Generation > Mission Specific Processing – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Additional processing steps not documented.
Basic	Additional processing steps documented. Additional processing steps not considered fit for stated purpose.
Good	Additional processing steps documented. All significant additional processing steps are fit for stated purpose.
Excellent	Additional processing steps documented. All additional processes steps considered fit for stated purpose.
Ideal	All additional processing steps are fully documented and considered state-of-the-art.

4.7. Detailed Validation

In this section we provide guidelines for the Detailed Validation assessment. The overall goal here is to verify that the mission performance is consistent with the sensor stated performance.

The detailed validation assessment is broadly divided into atmospheric column and geometric validation activities. Within these two sections are paired sub-sections describing each of the assessed performance metrics, each of which are evaluated both in terms of the quality of the validation method used and the validation results compliance. The results are reported as part of the Detailed Validation Maturity Matrix (5), which are then summarized across all performance metrics in the Validation Summary. This Validation Summary is the same summary presented in the Summary Product Evaluation Matrix shown in Figure 13 while the Detailed Validation Cal/Val Maturity Matrix is in Figure 16.

The remainder of this section includes:

- The criteria for grading the quality of the validation dataset, the validation method used, and validation completeness

- Assessment of the compliance of the product with the validation activity

- Each of the geometric performance metrics

- approach for synthesizing the results of the Detailed Validation into the Validation Summary is described.

Atmospheric Column				
Validation Summary		**Detailed Validation**		
Atmospheric Column Validation Methodology	←	Validation Dataset	Validation Method	Validation Completeness
Atmospheric Column Validation Results	←	Validation Results Compliance		
Geometric Validation	←	Sensor Spatial Response Method	Absolute Positional Accuracy Method	Temporal Stability Method
Geometric Validation Results	←	Sensor Spatial Response Compliance	Absolute Positional Accuracy Compliance	Temporal Stability Compliance

Figure 16. Detailed Validation Cal/Val Maturity Matrix and Validation Summary.

4.8. Validation Methodology

This section describes how, in generic terms, the criteria for grading the quality of the validation methodology, including the technique used, the validation approach (how mature and state-of-the-art the method is), and the completeness of the validation.

4.8.1. Validation Dataset

Generally, satellite validation attempts to demonstrate the compliance of data products with respect to some claimed performance level (e.g., documented specifications) by comparison of the product data with independent reference data.

The validation dataset section assesses the validation observations and suitability of the reference dataset for validation of these atmospheric column satellite data. The validation dataset should ideally be fully representative of the spatiotemporal variability of the satellite measurement. Any spatiotemporal or technique mismatch between validation and satellite data should be accounted for through an appropriate error analysis (e.g., root-mean-square difference relative to calculated uncertainties) and/or minimized wherever possible. Table 5-11 shows how the validation data are graded. The specific interpretation of these criteria in the quality assessment of a particular validation activity depends on a number of factors, therefore some level of expert judgement may be required when determining the grading.

Table 11. Validation > Validation Dataset – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Limited suitability of technique/dataset for satellite data validation
Good	Validation data is suitable for validation of satellite data, but no accounting for potential mismatch uncertainties.
Excellent	Validation data is suitable for validation of satellite data and technique/spatiotemporal mismatches are fully considered.
Ideal	Validation data is suitable for validation of satellite data. Technique/spatiotemporal mismatches are fully considered and related uncertainties are included in the uncertainty budget.

4.8.2. Validation Method

This section assesses the approach to the validation itself. Higher assessment grades will involve validation methods that are state-of-the-art, mature, and have a proven track record for validating atmospheric satellite data.

For higher grades, validation approaches will attempt to verify both the satellite measurements and their associated uncertainties. Validated uncertainties provide evidence of the credibility of the uncertainty estimate given. Commonly used metrics such as the statistical spread of differences may be used to estimate the uncertainty, however this often may not provide a realistic estimate of the actual uncertainty. Ideally, calculated uncertainties using first principals match the spread of comparisons between satellite and validation data sets as this means that the forward model assumptions (e.g., ray tracing, spectroscopy, instrument calibration) are robust.

In the same way, these guidelines describe how to assess the quality of satellite mission data. Similar considerations must be made for the quality of reference data used to validate the satellite mission data. The highest quality validation reference data have an associated uncertainty assessment and are traceable to the SI.

Table 12 shows how the validation approach is graded within the assessment framework.

Table 12. Validation > Validation Method – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Basic/outdated validation method, simple approach to uncertainty estimation from validation (i.e. spread of points around the fit). No quality information for validation reference dataset
Good	Mature validation approach with proven track-record, simple approach to uncertainty estimation from validation, good quality validation reference dataset with some uncertainty budgeting. Validation in line with NASA data readiness Stage 1 (Appendix A.2)
Excellent	Mature validation approach that is considered state-of-the-art. More sophisticated approach to uncertainty estimation from validation (e.g. includes satellite retrieval and validation method uncertainties). Excellent quality validation reference dataset with comprehensive uncertainty budgeting. Validation in line with NASA data readiness Stage 2 (Appendix B)
Ideal	Mature validation approach that is considered state-of-the-art. Metrologically robust approach to uncertainty estimation from validation. Excellent quality validation reference dataset with comprehensive uncertainty budgeting traceable to SI. Validation of data product and uncertainties in line with NASA data readiness Stage 3/4 (see Appendix B)

4.8.3. Validation Completeness

For spatiotemporally accurate and complete validation of atmospheric satellite data, validation activities must represent the full extent of measurements the satellite may make (e.g., global coverage, multi-year datasets, seasonal variability). This requires the use of a variety of reference datasets that cover different observation conditions.

This section assesses whether the validation methodology as a whole is representative of the entire range of scenarios that may reasonably be encountered during (e.g., northern and southern hemispheric observations, multi-year datasets, multi-season, variable albedo, and surface heights). The highest assessment grades will require validation across a range of these conditions.

Table 13 shows how the validation completeness is graded within the assessment framework.

Table 13. Validation > Validation Completeness – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Limited validation completeness, e.g. one single validation activity in space and/or time
Good	Multiple validation activities carried out over space and/or time. Allowance for some gaps in spatial/temporal coverage

Excellent	Multiple validation activities carried out over space and time. Intra-year temporal coverage (allowing for seasonality characterisation) and appropriate spatial coverage.
Ideal	Multiple validation activities carried out over space and time. Intra-year temporal coverage (allowing for seasonality characterisation) and appropriate spatial coverage. Assessment of uncertainties between validation sites or between validation activities at a given site.

4.9. Validation Results Compliance

This section assesses the results of the validation activities themselves. In the best-case scenario, these results will show that both the validated satellite measurements and their associated uncertainties have been obtained independent of the satellite data provider.

Grading for this subsection is based on the compliance of the validation results with current validation methods.

Table 14 shows how the validation results are graded within the assessment framework.

Table 14. Validation > Validation Compliance – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Claimed mission performance shows some agreement with validation results.
Good	Claimed mission performance shows good agreement with validation results.
Excellent	Claimed mission performance shows excellent agreement with validation results. Analysis performed independently of the satellite mission owner.
Ideal	Claimed mission performance shows excellent agreement with validation results, measurement uncertainties also validated. Analysis performed independently of the satellite mission owner.

4.10. Geometric Validation

There are three main aspects of assessing geometric performance in remote sensing data: 1) instrument sensor spatial response (SSR); 2) geolocation accuracy on the Earth's surface, or absolute positional accuracy (APA); and 3) multispectral sensor band-to-band registration (BBR). In geometric assessment, it is also important to consider temporal stability and global consistency in all aspects.

For geometric assessment, it is important whether the data are provided in a swath or gridded format. Swath data products have not been resampled and have the original time-tagged observations as sampled by the instrument. Gridded products typically contain observations that have been resampled to a fixed Earth grid with a fixed pixel interval and may be orthorectified to correct for terrain distortions.

Swath products must be accompanied by additional information regarding geometry of the observations in the product, either within the product or as a separate geolocation product.

This additional information usually includes time-tagged geodetic latitude and longitude of each observation (sample or pixel), and for many data sets, the terrain height. It may also include information such as the solar zenith and azimuth angles, quality flags, satellite position and its velocity and attitude, and the satellite zenith and azimuth angles. This data may be available for each observation or at a coarser resolution, e.g., at the scene center. For multispectral instruments there may be additional information about relative alignment of the individual bands, such as the band-to-band offsets.

For geometric validation of atmospheric column data, we consider the following metrics used for evaluation:

- Sensor spatial response (SSR)

- Absolute positional accuracy (APA)

- Multispectral sensor band-to-band registration (BBR)

- Temporal stability

These are each described in turn below, except for BBR, which is not relevant for atmospheric column measurements.

4.10.1. Sensor Spatial Response (SSR)

A sensor or detector spatial response is a function describing overall system response to a point impulse that is spatially located at every possible position. This spatial response function is called the system point spread function (PSF). A PSF is a spatial weighting function describing the responsivity of a detector to energy from a scene. A PSF may be constructed by two orthogonal line spread functions (LSFs), one in the along-track direction and another in the cross-track direction, for either a pushbroom, whiskbroom, or frame sensor instrument. A PSF is usually tested and analysed pre-launch and verified on-orbit. For gridded images, an LSF may be constructed in a cross-row or cross-column direction. Alternatively, an LSF may be derived from an edge spread function (ESF), which can be constructed from an image over a natural or man-made sharp edge feature. From the LSF, we can determine image quality parameters such as the footprint size at the full width at half maximum (FWHM), and the modulation transfer function (MTF). Alternatively, from an ESF, relative edge response (RER) can be determined as an image quality parameter. In general, we want the MTF to be at least 0.25 or greater at the Nyquist frequency (one cycle every per two times the ground sample distance). Note that for gridded products, the MTF can be improved by aggregating or downsampling the data at a larger pixel size. For multispectral instruments, these measurements should be made separately for each spectral band. Also, the spatial response may vary by position within the focal plane, e.g., by detector, so measurements should be made to understand any detector-specific variation that may be present.

4.10.2. Absolute Positional Accuracy (APA)

As agency and commercial satellite sensors become more advanced and numerous, with many providing high resolution or very high resolution (VHR) imagery, it is important to evaluate the positional accuracy of the products against the accuracy specifications and typical user needs.

Geolocation accuracy assessment typically involves evaluation of the positional accuracy of the data using ground truth with a known geolocation accuracy, typically ground control points (GCPs). For many applications, the geolocation accuracy should have a circular error at the 90th percentile (CE90) to within 0.5 of the product pixel size for gridded products, and within 0.5 of the ground sample distance for swath products, or within 0.5 of the sensor's footprint size measured at the full width at half maximum (FWHM) of its PSFs if that is available. The GCPs should be as evenly distributed geographically as possible, to ensure consistency in the geolocation accuracy assessment globally. For sensors with numerous detectors acquiring data simultaneously, to ensure an unbiased assessment due to image distortion, GCPs should be evenly distributed over the entire detector array.

For swath data, the accompanying geolocation information in the geolocation product is used to compare the geolocated observations to the ground truth. Note that for multi-spectral data, the geolocation accuracy may be assessed using a single band, but may also be done for individual bands, and so may be impacted by band-to-band registration.

Should the data in a single scene be used for object identification, for example, a geolocation error of a few pixels may not be significant, and thus further geolocation error correction may not be required for the application. However, should the data be used for time series analyses, these same geolocation errors will result in unusable data for this purpose. Relative geolocation errors could be reduced by aggregating or down sampling the data to a larger pixel size.

4.10.3. Temporal Stability

Because of potential long-term changes in sensor characteristics, it is necessary to monitor an instrument's performance over the entire mission to ensure that any changes in performance over time are understood. The validation stages defined by the CEOS Land Product Validation subgroup include requirements for spatial and temporal consistency. This consistency cannot be assessed without adequate geometric temporal stability.

Ideally, the satellite data products are evaluated over globally representative locations. Absolute positional accuracy methods can be used to quantify the positional stability of sensor products, and these can be applied multiple times over a season and/or years to assess the temporal stability of satellite data products.

It is a challenge to achieve sub-pixel accuracy for images at very high resolution. It is also recognized that there is not an overabundance of globally distributed points of absolute ground truth. High resolution or VHR images are often used as reference for calibration and validation of geolocation performance, but caution should be used, as the uncertainties of these reference images can exceed the pixel size of VHR images. Users of EO data are often require

temporal stability at particular sites for time series analyses and thus temporal stability is an important aspect of geolocation accuracy.

4.11. Validation Summary

The validation summary provides a synthesis of the per-performance metric assessments provided in the Detailed Validation Cal/Val Maturity Matrix (Figure 16). It is also presented as part of the Summary Cal/Val Maturity Matrix.

Each row in the Detailed Validation Cal/Val Maturity Matrix is represented by one cell in the Validation Summary column. Thus, there are two summary cells in total – Atmospheric Column Validation Methodology and Atmospheric Column Validation Compliance The grade for each of these summary cells represents a combination of the grades of the contributing cells. The approach is to effectively average the grades of the contributing cells, where each grade is valued as follows: Basic is 1, Good is 2, Excellent is 3, and Ideal is 4.

4.12. Atmospheric Column Product Overall Grade

Using the detailed criteria from the previous sections as a guide, an overall grade of the product should be provided to guide the user of data in its utility for science or policy or applications.

Table 15. Atmospheric Column Product Overall Grade.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Product is not assessable
Basic	Products have demonstrated skill in quantifying "facility scale" column amounts; however, there is insufficient documentation, VVUQ, reproducibility and traceability for these data to be effectively used for decision making purposes.
Good	Products can be used for corroboration purposes only and not for independent analysis. Reported products have limited documentation, VVUQ, reproducibility and traceability.
Excellent	Products can be independently used for science analysis or applications or decision making. However, there may be incomplete product description or detailed validation
Ideal	All aspects of the quality assessment are ideal and meet best practices. Reported products are traceable to L0 / L1.

5. Overview: Quality Assessment for Reporting Methane EmissionsEmissions

In recent years, the increasing range of applications of earth observation (EO) data products and availability of low-cost satellites has resulted in a growing number of commercial EO satellite systems, developed with a view to deliver end-to-end information services, many of which sense the atmospheric domain. This evolution in the marketplace has led to increasing interest from space agencies in the acquisition of commercial EO data products, as they may provide complementary capabilities and services to those they currently offer.

To ensure that decisions on commercial data acquisitions can be made fairly and with confidence, there is a need for an objective framework with which their data quality may be assessed. The ESA Earthnet Data Assessment Pilot (EDAP) project therefore defines this EO mission quality assessment framework, within which the project performs quality assessments of commercial satellite missions in the optical, SAR and atmospheric domains. Presented here is the latest evolution of this framework for atmospheric missions that provide measurements of greenhouse gas (GHG) column enhancements at facility scale (~10 meters to 100 meters) and corresponding estimates of emissions from these enhancements. In particular this document focuses on emission estimates from these measurements. The previous section focuses on the measurement of column enhancements.

This section is intended to provide specific guidelines for mission quality assessment of atmospheric sensors, specifically for emission products (Level 4) derived from atmospheric trace gas column data (Level 2) as part of the implementation of the generic EO mission quality assessment (Hunt, 2021) for this domain. Our quality assessment includes (1) traceability of the data to the SI through known standards, (2) reproducibility of emission estimates given the reported products and documentation, (3) transparency of the emission estimates, and (4) VVUQ (Validation, Verification, Uncertainty Quantification) of the emissions estimates. Section 5.2 provides a summary of the mission quality assessment framework. Section 5.3 provides a review of the atmospheric mission quality, as evidenced by its documentation. Section 5.4 provides guidelines for verifying the mission data quality is consistent with its stated performance. Section 5.5 describes the overall quality assessment guidelines that should result from the quality assessment from the previous sections.

5.1. EO Mission Quality Assessment Framework Summary

This section outlines the EO mission quality assessment for emission products. The evaluation is primarily aimed at verifying that mission data has achieved the claimed mission performance and, where applicable, reviews the extent to which the missions follow community best practice in a manner that is "fit for purpose".

The approach taken to assess data product quality is based on the QA4EO principle (QA4EO Task Team, 2010) and builds on the structure and reporting style developed in other similar work (e.g., Nightengale et al., 2019). This quality assessment framework, developed within the ESA Earthnet Data Assessment Pilot (EDAP) project, aims to build on the experience of this previous work targeting the satellite Cal/Val context.

The assessment itself is conducted in two parts, as follows:

- Documentation Review – review of mission quality as evidenced by its documentation.

- Detailed Validation – quantitative assessment of product compliance with stated performance.

These parts of the assessment, along with their grading criteria, are described in Sections 3 and 4, respectively. The activities are divided into sections and subsections constituting each of the different aspects of data product quality that are assessed and graded. Assessment results are provided in a separate Quality Assessment (QA) Report and are also summarized in a color-coded Product Evaluation Matrix.

It is expected that all relevant mission information needed to perform the assessment would be available to all users, however it is understood that confidentiality may be required for some aspects of a mission. Where this is the case, it will be indicated as confidential in the quality assessment report. In general, pertinent key conclusions of confidential documentation should nevertheless be published openly to the extent possible.

5.2. Mission Data Supply Chain Assessment Overview

The product assessment outlined in this document forms part of a wider supply chain assessment summary (Figure 13). This overview matrix encompasses documentation review and detailed validation assessments for all data processing steps for a given atmospheric mission, including calibrated radiances (Level 1B), retrieved atmospheric column products (Level 2), and further derived emission (Level 4), if applicable.

For each of the rows of the Supply Chain Summary in Figure 12, a full set of data product quality assessment guidelines exist in the same format as shown in this document for the emission products.

To ensure a complete and transparent quality assessment, EO missions yielding an emission data product must also include a documented and detailed validation assessment for the associated L1B calibrated radiances and the L2 column products used to obtain the emission in order to be fully assessed under EDAP guidelines. Ideally, the reported L4 emission estimates are reproducible from the L1 and L2 products and associated documentation (e.g., ATBDs and product descriptions).

5.3. Quality Assessment Report

The quality assessment for a given emission product is reported using the QA Report template. The template ensures consistency of reporting and facilitates comparison between the assessments of similar missions. The QA Report covers each section of analysis, providing more detailed information, and a completed mission product evaluation matrix (see following subsection) presenting the results of each sub-section of analysis in a color-coded table.

5.4. Product Evaluation Matrix

The product evaluation matrix provides a high-level color-coded summary of the quality assessment results. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. Subsection grades are indicated by the color of the respective grid cell, which are defined in the key. A padlock symbol in the corner of given cell indicates that the information used to assess the respective subsection is not available to the public. The reporting of assessment results is divided between two evaluation matrices, as follows:

- Summary Product Evaluation Matrix
- Detailed Validation Maturity Matrix

These matrices are described below.

5.4.1. Summary Product Evaluation Matrix

The Summary Product Evaluation Matrix is shown in Figure 17. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. The matrix on the left (in dark blue) summarizes the results of the Documentation Review, while the additional column on the right (in light blue) summarizes the results of the Detailed Validation. The Validation Summary column is separated from the main table to make clear the results can come from multiple assessment sources.

Data Provider Documentation Review				Validation Summary		Key
Product Information	Metrology	Product Generation				Not Assessed
Product Details	Metrological Traceability Documentation	Emission Quantification Method		Emission Validation Methodology		Not Assessable
						Basic
Availability & Accessibility	Uncertainty Characterization	Mission Specific Processing		Emission Validation Results		Good
						Excellent
Product Format, Flags & Metadata	Ancillary Data					Ideal
User Documentation						🔒 Not Public

Figure 17. Summary Product Evaluation Matrix.

5.4.2. Detailed Validation Maturity Matrix

The Detailed Validation Maturity Matrix (Figure 18) provides more complete reporting of analysis contributing to the Validation Summary – breaking down the validation methodologies used and the results. This section is aimed at the more technically focused reader. Since, for a given mission, multiple validation studies may be performed – for example, by the mission/vendor and/or by independent assessors – there can be multiple Detailed Validation Maturity Matrices produced and reported. Detailed evaluation (right side) should be performed first, and the grades used generate the validation summary (left side).

Emission Validation			
Validation Summary	Detailed Validation		
Emission Validation Methodology	←	Validation Technique	Validation Approach	Validation Completeness
Emission Validation Results	←	Validation Results Compliance		

Key
Not Assessed
Not Assessable
Basic
Good
Excellent
Ideal

🔒   Not Public

Figure 18. Validation Maturity Matrix, showing the Validation Summary column from the Product Evaluation Matrix.

5.5. Approach to Grading

The assessment framework is aimed at verifying the claimed mission performance, and to assure that the mission follows community best practice to an extent that is "fit for purpose". The grading criteria for each category are determined based on a logical interpretation of this principle. For example, pre-launch calibration quality grading is based on the comprehensiveness of activity with respect to the target instrument performance.

Grades of Basic, Good, Excellent, or Ideal may be given. The Ideal grade level is generally reserved to provide recognition for achieving the highest standard of quality with respect to community best practice. This high bar of quality may be aspirational but is the benchmark that EO data providers should aim for. Note that a grade of Basic can be considered acceptable in a given context. The criteria for grading each box of the matrix are described in Sections 3 and 4.

Additionally, a subsection may also indicate Not Assessable or Not Assessed. These cover the cases where certain aspects of product quality will not be assessed – either because there is insufficient information available to make an assessment, or because it is out of scope of the assessment.

5.6. Data Provider Documentation Review

In this section we provide detailed guidelines for *Data Provider Documentation Review*. This assessment aims to review mission quality as evidenced by its documentation. It is divided into the follow sections:

- Product Information
- Metrology
- Product Generation

In the following we look at each of these sections in turn and discuss the grading criteria.

The results of the Documentation Review are reported on the left portion of the Summary Product Evaluation Matrix (Figure 17). This portion is shown in Figure 19.

Data Provider Documentation Review		
Product Information	**Metrology**	**Product Generation**
Product Details	Metrological Traceability Documentation	Emission Quantification Method
Availability & Accessibility	Uncertainty Characterization	Mission Specific Processing
Product Format, Flags & Metadata	Ancillary Data	
User Documentation		

Figure 19. Data Provider Documentation Review Matrix.

5.7. Product Information

The Product Information section covers the top-level product descriptive information, product format, and the supporting documentation. Its subsections are now defined.

5.7.1. Product Details

Certain basic descriptive information should be provided with any EO data product and is required for assessment of all mission domains. The list of this required information is as follows:

- Product name
- Sensor Name

- Sensor Type: Describe sensor design type, e.g., multi-channel, hyperspectral, interferometer etc., and spectral domains, e.g., visible (VIS), near infrared (NIR), shortwave infrared (SWIR), thermal infrared (TIR).

- Mission Type: Either single satellite or constellation of a given number of satellites.

- Mission Orbit: For example, Sun Synchronous Orbit with Local Solar Time.

- Product version number

- Product ID

- Processing level of product

- Spatial coverage

- Point of contact (Responsible organization, including email address)

- Product access (e.g., URL, DOI if applicable)

- Restrictions for access and use, if any

Table 16 shows how provision of data product information relates to its grade for this sub-section of the quality assessment.

Table 16. Product Information > Product Details – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside of the scope of study.
Not Assessable	Relevant information not made available.
Basic	Many pieces of important information are missing.
Good	Some pieces of important information are missing.
Excellent	Almost all required information available.
Ideal	All required information available.

5.7.2. Availability & Accessibility

This is about how readily the data are available to those who wish to use them. It does not necessarily require cost-free access but is more about following the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles for scientific data management and stewardship (Wilkinson et al., 2016), which provide valuable principles for all data applications. These state that:

- Data should be **findable**:
  - o Metadata and data are assigned a globally unique and persistent identifier
  - o Data are described with rich metadata
  - o Metadata clearly and explicitly include the identifier of the data it describes
  - o Metadata and data are registered or indexed in a searchable resource

- Data should be **accessible:**

  o Metadata and data are retrievable by their identifier using a standardised communications protocol

  o The protocol is open, free and universally implementable

  o The protocol allows for an authentication and authorisation procedure where necessary

- Data should be **interoperable**:

  o Metadata and data use a formal, accessible, shared and broadly applicable language for knowledge representation

  o Metadata and data use vocabularies that themselves follow FAIR principles

  o Metadata and data include qualified references to other (meta)data

- Data should be **reusable**:

  o Metadata and data are richly described with a plurality of accurate and relevant attributes

  o Metadata and data are released with a clear and accessible data usage license

  o Metadata and data are associated with detailed provenance

  o Metadata and data meet domain-relevant community standards

Table 17 shows how a data product's provision of the above information relates to the grade it achieves for this sub-section of the quality assessment.

Table 17. Product Information > Availability and Accessibility – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Relevant information not made available.
Basic	The data set does not appear to be following the FAIR principles
Good	The data set meets many of the FAIR principles and/or there is an associated data management plan that shows progress towards the FAIR principles
Excellent	The data set meets many of the FAIR principles and has an associated data management plan and is available either free of cost or through an easy-to-access commercial license.
Ideal	The data set fully meets the FAIR principles and has an associated data management plan and is available either free of cost or through an easy-to-access commercial license.

5.7.3. Product Format, Flags and Metadata

An important aspect of EO data products that ensures ease of access to the widest variety of users is their format. Product metadata and flags offer users important extra layers of useful

descriptive information, in addition to the measurements themselves, that can be crucial to their analysis.

In the ideal case, the product format would meet the appropriate Committee on Earth Observation Satellites (CEOS)-Analysis Ready Data (ARD) metadata guidelines (CEOS LSI, 2020) requirements.

In the case where such a standard does not exist, product format is graded based on the following:

- the extent to which it is documented

- whether a standard file format is used (e.g., NetCDF)

- whether it complies with standard variable, flag and metadata naming conventions, such as the Climate and Forecast (CF) metadata Conventions (Eaton et al., 2020), or, for data from the European Union, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive (INSPIRE Drafting Team, 2013)

- whether flags and metadata provide an appropriate breadth of information

If product is derived from a constellation of satellites, the specific satellite used should be included in the product metadata.

Table 18 shows how a given EO data product should be graded for its format.

Table 18. Product Information > Product Format, Flags and Metadata – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Non-standard, undocumented data format.
Basic	Non-standard or proprietary data format, or poorly documented standard file format. Minimal useful metadata or data flags provided.
Good	Data exist in a documented standard file format. Non-standard naming conventions used. Includes a good set of documented metadata and data flags.
Excellent	Data are organized in a well-documented standard file format, meeting community naming convention standards. Comprehensive set of metadata and data flags.
Ideal	Analysis Ready Data standard if applicable, else as *Excellent*.

5.8. User Documentation

Data products should be accompanied with the following minimum set of documentation for users, which should be regularly updated as required:

- Product User Guide/Manual (PUG/PUM)

- Algorithm Theoretical Basis Document (ATBD)

It may be for a given mission that in place of these documents some combination of articles, publications, webpages and presentations provide a similar set of information. For the highest grades however, they should be presented as a formal document, since users should not be expected to search the information out. The QA4ECV project provides guidance for the expected contents of these documents (INSPIRE Drafting Team, 2013), which they can be evaluated against.

Table 19 describes how the assessment framework grades a products user documentation.

Table 19. Product Information > User Documentation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No user documentation provided or documentation out-of-date.
Basic	Limited PUG available, no ATBD. Information is up-to-date.
Good	Some PUG and ATBD-type information available. These may be formal documents or from multiple sources. Documentation is up-to-date.
Excellent	PUG meets QA4ECV standard, reasonable ATBD. Documents are up-to-date.
Ideal	PUG and ATBD available meeting QA4ECV standard. Documents are up-to-date.

5.9. Metrology

Metrology is the science of measurement. This section covers the aspects of the mission-related to measurement quality, including calibration, traceability and uncertainty. The Metrology subsections are now defined.

5.9.1. Metrological Traceability Documentation

Traceability is defined in the vocabulary of metrology (VIM) (JCGM, 2012) as, "property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty" and reinforced in the QA4EO procedures. Traceability is therefore a key aspect of achieving reliable, defensible measurements. In this definition an important part of measurement traceability is highlighted – that it is well documented. This of course must be the case for EO data products too.

Various diagrammatic approaches have been developed to present the traceability chains for EO data products (e.g. the QA4ECV guidance, which includes a traceability chain drawing tool (Scanlon, 2017)). Such a diagram should be included in the documentation for every EO mission. The FIDUCEO project has provided guidance for a more detailed measurement function centered "uncertainty tree diagram" which is ultimately more suitable for most examples of EO data processing and should be the aspiration for missions in the future (Burgdorf et al, 2019).

Table 20 shows how the assessment framework grades the metrological traceability documentation, based on its completeness.

Table 20. Metrology > Metrological Traceability Documentation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No traceability chain documented.
Basic	Traceability chain diagram and/or uncertainty tree diagram included, missing some important steps.
Good	Traceability chain and/or uncertainty tree diagram documented identifying most important steps and sources of uncertainty.
Excellent	Rigorous uncertainty tree diagram, with a traceability chain documented, identifying all reasonable steps and accompanying sources of uncertainty.
Ideal	Rigorous uncertainty tree diagram and traceability chain documented, identifying all reasonable steps and accompanying sources of uncertainty. Establishes traceability to SI.

5.9.2. Uncertainty Characterization

To ensure measurements are both meaningful and defensible, it is crucial that they include rigorously evaluated uncertainty estimates. A comprehensive description of how to evaluate sources of uncertainty in a measurement, and propagate them to a total uncertainty of the final measurand, is provided by the metrological community in the Guide to the Expression of Uncertainty in Measurement (GUM) (JCGM, 2008).

The application of Earth Observation metrology has progressed greatly in recent years. Increasingly, providers of operational and reprocessed data products are applying different approaches to evaluate and distribute metrologically rigorous error-covariance at the per pixel level, as required by climate studies. For example, ESA's Sentinel-2 mission has developed an on-the-fly, pixel-level uncertainty evaluation tool (Fox, 2019). There have also been some initiatives, like the previously mentioned FIDUCEO project, that have applied metrology to historical sensor data records (Vendt, 2020).

With that said, it is typical for uncertainties (or performance estimates) to be evaluated in a manner that does not comply with the GUM. With that said, it is typical for uncertainties (or performance estimates) to be evaluated in a manner that does not comply with the GUM, for example, the relative offset from a comparison emission observation may be quoted as the uncertainty. Higher grades should attempt to quantify and propagate all sources of uncertainty associated with emission quantification, e.g. wind speed, enhancement area, etc.

Table 21 shows the uncertainty characterization grading under the assessment framework.

Table 21. Metrology > Uncertainty Characterisation – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No uncertainty information provided.
Basic	Uncertainty established by limited comparison to measurements by other sensor/s.
Good	Limited use of rigorous uncertainty estimation approaches, and/or, an expanded comparison to measurements by other sensors. Most important sources of uncertainty are included.
Excellent	Metrologically rigorous approach used to estimate measurement uncertainty, all important sources of uncertainty are included. Uncertainty per pixel provided.
Ideal	Metrologically rigorous approach used to estimate measurement uncertainty, including a treatment of error-covariance. Per pixel uncertainties in components, e.g., random systematic – as appropriate for the error-correlation structure of the data.

5.9.3. Ancillary Data

Throughout the processing chain there may be a requirement for external input data, for example, a priori atmospheric state information, or reference data for algorithm tuning. The ancillary datasets used during the processing should be identified to the user (where possible due to commercial sensitivity). Ideally this should be traceable on a per product level.

Ancillary datasets must be of a sufficient quality, including the application of suitably rigorous metrology, for example, in the form of SI traceability.

The suitability of the ancillary data for its application must also be considered, with respect to the mission's stated performance requirements. For example, the quality, size, and representativeness of algorithm input data. The requirements will be specific to the retrieval method used and may require some expert judgement.

Table 22 shows how the ancillary data are graded under the assessment framework.

Table 22. Metrology > Metrology > Ancillary Data – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Use of ancillary data undocumented.
Basic	Ancillary data used in product generation, specified to some extent, though incomplete. Not entirely of a sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.
Good	Ancillary data used in product generation, specified, though not necessarily on a per product basis. Mostly of a sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.
Excellent	Ancillary data used in product generation, fully specified per product, and traceable. Ancillary data used are of sufficient quality to be judged "fit for purpose" in terms of the mission's stated performance.

| Ideal | Ancillary data used in product generation, meets the Excellent criteria, and are traceable to SI where appropriate. |

5.10. Product Generation

The Product Generation section covers the processing steps undertaken to produce the data product. This primarily concerns the quantification of emissions from L2 atmospheric trace gas column data, and further post-processing steps that may be undertaken.

5.10.1. Emission Quantification Method

A multitude of emission quantification approaches exist that are suited to different emission source types. For example, the use of Integrated Mass Enhancement (IME) and cross-sectional emission techniques are well suited for point source emissions (Gorroño et al., 2023) where the entire emission plume can be resolved and isolated from background pixels. In contrast, estimation of surface emissions via inversion of satellite observations with a chemical transport model (constrained by prior emission inventory data) is best suited for more diffuse sources with a wider spatial extent.

The emission quantification method should be of a sufficient quality that it is "fit for purpose" within the context of the mission's stated performance across all stated use cases (e.g., scene types, emission source types). What this requires is specific to a given variable's retrieval methods and will require a degree of expert judgement.

Table 23 shows how the assessment framework grades the retrieval algorithm used to generate L2 products.

Table 23. Product Generation > Emission Quantification Method – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Emission quantification method not documented.
Basic	Emission quantification method somewhat documented. Emission quantification method either too simple or poorly suited to the target emission sources to be judged "fit for purpose" in terms of the mission's stated performance.
Good	Emission quantification method is well documented. Reasonable emission quantification method used, judged "fit for purpose" in terms of the mission's stated performance for most expected use cases, with at least a sensitivity analysis carried out.
Excellent	Emission quantification method is well documented and published via peer review. Emission quantification method "fit for purpose" in terms of the mission's stated performance for all expected use cases and validated performance against similar approaches or with empirical evidence.
Ideal	In addition to meeting the excellent criteria, the full uncertainty budget for the emission estimate are described including the uncertainties from the methane plume definition and the approach used to relate the plume enhancements to emissions.

5.10.2. Mission-Specific Processing

Additional processing steps are separate to the main retrieval processing. These may include processes like the generation of quality or cloud masks. Additional processing steps must themselves be assessed for quality based on their "fitness for purpose" in the context of the mission.

In the case of additional processes where the measurement data themselves are transformed in some manner, such as orthorectification, the uncertainties from the measurement data must be propagated, as well as introducing appropriate additional uncertainty components caused by the processing itself. This is required for the uncertainties to remain meaningful.

Each additional processing step should be separately assessed based on the criteria described in Table 24, and then a combined score determined.

Table 24. Product Generation > Mission Specific Processing – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Additional processing steps not documented.
Basic	Additional processing steps documented. Additional processing steps not considered fit for stated purpose.
Good	Additional processing steps documented. All significant additional processing steps are fit for stated purpose.
Excellent	Additional processing steps documented. All additional processes steps considered fit for stated purpose.
Ideal	All additional processing steps are fully documented and considered state-of-the-art.

5.11. Detailed Validation

In this section we provide guidelines for the *Detailed Validation* assessment. The overall goal here is to verify that the mission performance is consistent with the sensor stated performance.

The detailed validation assessment is broadly divided into the validation methodology, and the validation results compliance. Within these two sections are paired sub-sections describing each of the assessed performance metrics, each of which are evaluated both in terms of the quality of the validation method used and the validation results compliance. The results are reported as part of the Detailed Validation Maturity Matrix (Figure 20), which are then summarised across all performance metrics in the Validation Summary. This Validation Summary is the same summary presented in the Summary Product Evaluation Matrix shown in Figure 17.

The remainder of this section includes:

- The criteria for grading the quality of the validation methodology, including the validation dataset, method, and completeness.

- Assessment of the compliance of the product with the validation activity

- The approach for synthesizing the results of the Detailed Validation into the Validation Summary.

Emission Validation			
Validation Summary	**Detailed Validation**		
Emission Validation Methodology ←	Validation Technique	Validation Approach	Validation Completeness
Emission Validation Results ←	Validation Results Compliance		

Figure 20. Detailed Validation Cal/Val Maturity Matrix and Validation Summary.

5.11.1. Validation Methodology

This section describes how, in generic terms, the criteria for grading the quality of the validation data set, including the technique used, the validation approach (how mature and state-of-the-art the method is), and the completeness of the validation.

5.11.2. Validation Data Set

Generally, satellite validation attempts to demonstrate the compliance of data products with respect to some claimed performance level (e.g., documented specifications) by comparison of the product data with independent reference data. For satellite-derived emission data, the reference data usually takes the form of a controlled release of a known quantity of trace gas, although this assessment does not strictly limit validation activities to controlled release comparison experiments. Validation against emission estimates from other satellites will only be able to achieve lower assessment grades due to the lack of traceability of the reference dataset.

The validation technique section assesses the validation activity observations themselves and assesses both the description of the validation technique and suitability of the reference dataset for validation of atmospheric satellite data.

Table 25 shows how the validation technique is graded. The specific interpretation of these criteria in the quality assessment of a particular validation activity depends on a number of factors, therefore some level of expert judgement may be required when determining the grading.

Table 25. Validation > Validation Data – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.

| Not Assessable	No validation activity performed.
Basic	Limited suitability of dataset for satellite data validation.
Good	Full description of validation data, validation data is suitable for validation of satellite data, but no accounting for potential mismatch uncertainties.
Excellent	Validation data is suitable for validation of satellite data and technique mismatches are fully considered.
Ideal	Full description of validation technique, validation data is suitable for validation of satellite data. Data mismatches are fully considered and related uncertainties are included in the uncertainty budget.

5.11.3. Validation Method

This section assesses the approach to the validation itself. Higher assessment grades will involve validation methods that are state-of-the-art, mature and have a proven track record for validating atmospheric satellite data. For higher grades, validation approaches will attempt to verify both the satellite measurements and their associated uncertainties. Validated uncertainties provide evidence of the credibility of the uncertainty estimate given. Commonly used metrics such as the statistical spread of differences may be used to estimate the uncertainty, however this often may not provide a realistic estimate of the actual uncertainty.

In the same way, these guidelines describe how to assess the quality of satellite mission data. Similar considerations must be made for the quality of reference data used to validate the satellite mission data. For the particular case of emission validation techniques involving controlled releases, the quality of the "known" emission estimate used in comparison studies will be a primary assessment criterion. The uncertainty of the release estimate itself should ideally be fully budgeted, with all uncertainty contributions accounted for. SI-traceable controlled emissions (e.g. from the NPL Controlled Release Facility) are required for the highest assessment grades.

Table 26 shows how the validation approach is graded within the assessment framework.

Table 26. Validation > Validation Approach – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Basic/outdated validation method, simple approach to uncertainty estimation from validation (i.e. spread of points around the fit). No quality information for validation reference dataset
Good	Mature validation approach with proven track-record, simple approach to uncertainty estimation from validation, good quality validation reference dataset with some uncertainty budgeting. Validation in line with NASA data readiness Stage 1 (Appendix A.2)
Excellent	Mature validation approach that is considered state-of-the-art. More sophisticated approach to uncertainty estimation from validation (e.g. includes satellite retrieval and validation method uncertainties). Excellent quality validation reference dataset with comprehensive uncertainty budgeting. Validation in line with NASA data readiness Stage 2 (Appendix B)

	Mature validation approach that is considered state-of-the-art. Metrologically robust approach to uncertainty estimation from validation (e.g. includes both satellite emission and validation method uncertainties, considers error correlations). Excellent quality validation reference dataset with comprehensive uncertainty budgeting traceable to SI. Validation of data product and uncertainties in line with NASA data readiness Stage 3/4 (see Appendix B)
Ideal	

5.11.4. Validation Completeness

For accurate and complete validation of satellite emissions data, validation activities must cover the full extent of observations the satellite may make (e.g., range of windspeeds and emission rates, range of surface biomes/surface reflectance). This may require the use of a variety of different reference datasets to cover different observation conditions.

This section assesses that the validation methodology as a whole covers the entire range of scenarios that may reasonably be encountered during a given retrieval scene. Satellite emission validation activities are often carried out as individual case studies, and the network-based validation approach of L1B or L2 atmospheric products is not shared with L4 emission data. However, the highest assessment grades should aim to characterize a range of emission observation scenarios. Additionally, studies where multiple teams have carried out independent emission quantification for the same satellite data as part of a validation exercise will also achieve higher grades.

Table 27 shows how the validation completeness is graded within the assessment framework.

Table 27. Validation > Validation Completeness – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Limited validation completeness, e.g. one single validation datapoint.
Good	Some coverage of different emission scenarios within validation efforts (e.g. differing emission rate/windspeed).
Excellent	Good coverage of different emission scenarios within validation efforts (e.g. differing emission rate/windspeed). Validation activity may involve multiple reference emission datasets encompassing different scene types, or multiple independent analyses of the same satellite dataset.
Ideal	Excellent coverage of different emission scenarios within validation efforts (e.g. differing emission rate/windspeed). Validation activity will involve multiple reference emission datasets encompassing different scene types, or multiple independent analyses of the same satellite dataset.

5.11.5. Validation Results Compliance

This section assesses the results of the validation activities themselves. In the best-case scenario, these results will show that both the validated satellite measurements and their associated uncertainties have been obtained independent of the satellite data provider.

Grading for this subsection is based on the compliance of the validation results with current validation methods.

Table 28 shows how the validation results are graded within the assessment framework.

Table 28. Validation > Validation Compliance – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	No validation activity performed.
Basic	Claimed mission performance shows some agreement with validation results.
Good	Claimed mission performance shows good agreement with validation results.
Excellent	Claimed mission performance shows excellent agreement with validation results. Analysis performed independently of the satellite mission owner.
Ideal	Claimed mission performance shows excellent agreement with validation results, measurement uncertainties also validated. Analysis performed independently of the satellite mission owner.

5.11.6. Validation Summary

The Validation Summary provides a synthesis of the per performance metric assessments provided in the Detailed Validation Cal/Val Maturity Matrix (Figure 20). It is also presented as part of the Summary Cal/Val Maturity Matrix. Each row in the Detailed Validation Cal/Val Maturity Matrix is represented by one cell in the Validation Summary column. Thus, there are two summary cells in total – Emission Validation Methodology and Emission Validation Compliance. The grade for each of these summary cells represents a combination of the grades of the contributing cells. The approach is to effectively average the grades of the contributing cells, where each grade is valued as follows: Basic is 1, Good is 2, Excellent is 3, and Ideal is 4.

5.12. Emission Product Overall Grade

Using the detailed criteria from the previous sections as a guide, an overall grade of the product should be provided to guide the user of data in its utility for science or policy or applications.

Table 29. Emission Product Overall Grade – Assessment Criteria.

Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Product is not assessable
Basic	Products have demonstrated skill in quantifying "facility scale" emissions; however, there is insufficient documentation, VVUQ, reproducibility and traceability for these data to be effectively used for decision making purposes.
Good	Products can be used for corroboration purposes. Reported products have limited documentation, VVUQ, reproducibility and traceability.
Excellent	Products (emissions) can be independently used for science analysis or applications or decision making. However, there may be incomplete product description or detailed validation

Ideal	All aspects of the quality assessment are ideal and meet best practices. Reported products are traceable to L0 / L1.

Appendix A. Validation Methods for Atmospheric Column Products

This appendix offers a short summary of some methods for retrieved atmospheric column data validation.

Atmospheric column data retrieved by satellites are typically validated (and often bias corrected) via direct comparison with ground-based remotely sensed atmospheric column data from fixed sites, or via comparison against in situ observations made throughout a given atmospheric profile.

The following sections of this appendix each describe a commonly used validation method, by specifying the following:

- Description – general outline of method, with appropriate references.

- Scope of Representativeness – Comparability of validation data/method with satellite data/method, as well as the spatiotemporal extent and maturity of validation method,

- Quality – best uncertainty achievable with this method, according to literature.

A.1. Ground-Based Methods

Validation of trace gas column satellite products is often carried out via intercomparison with ground-based networks of Fourier Transform Infrared (FTIR) spectrometers or in situ profiles from (for example) aircraft or balloon, as discussed in subsequent sections. Validation (and bias correction) against these networks is often carried out automatically as part of the retrieval processing chain, and validation is carried out upon each satellite overpass of a ground-based network site.

A.1.1. Total Carbon Column Observing Network (TCCON)

Description

TCCON has been a longstanding tool for validating satellite GHG column data, such as CO2 column products from GOSAT, GOSAT-2 and OCO-2, and CH_4 column from the Sentinel-5P TROPOMI instrument. The network consists of 23 Bruker IFS 125HR FTIR spectrometers, with a spectral resolution of $\approx$-.02 cm^{-1}. These instruments retrieve total column amounts of CO2, CH4, N2O, CO, and HDO from direct solar observations in the near-infrared (Wunch et al., 2011).

Column retrievals from TCCON sites have themselves been calibrated using aircraft and balloon-borne in situ observations and are therefore traceable to WMO in situ GHG calibration standards. TCCON validation forms a key traceability link between in situ and satellite GHG observations (Wunch et al., 2010, Messerschmidt et al., 2011).

Scope of Representativeness

Considerations include:

- Directly compatible satellite and validation data products (both full atmospheric column products retrieved from radiometric observations)

- Relatively wide spatial distribution of TCCON sites (albeit some gaps in coverage)

- Long-term continuous dataset (>10 years)

Quality

Quality expectations are variable, depending on individual validation activity (site, satellite, and time dependent). Example 1σ in situ instrument precisions: e.g., <1 µmol·mol^{-1} for CO_2, <5 nmol·mol^{-1} for CH_4 (1σ) (Wunch et al., 2011).

A.1.2. Collaborative Carbon Column Observing Network (COCCON)

Description

A key issue with the established TCCON validation network is the uneven distribution of sites and hence limited spatial coverage in certain regions (Africa, South America, and parts of Asia in particular) (Wunch et al., 2017). COCCON is designed to supplement the existing TCCON network and remedy the shortcomings of TCCON. COCCON consists of Bruker EM27/SUN model FTIR solar absorption spectrometers, which share the same concept of operation as the TCCON instruments.

The key difference between the network instruments is that the EM27/SUN model is portable, easy to deploy, and lower cost than the fixed TCCON instruments. More COCCON instruments can therefore be deployed, and these can be selectively distributed in order to fill the spatial gaps of the TCCON network.

Long-term performance of COCCON instruments have been assessed against existing TCCON instrumentation, showing good agreement and stability over a period of several years. Additionally, the use of an EM27/SUN travelling standard instrument has been proposed to ensure close TCCON-COCCON calibration and to link COCCON to the WMO traceability chain (Frey et al., 2019).

Scope of Representativeness

Considerations include:

- Directly compatible satellite and validation data products (both full atmospheric column products retrieved from radiometric observations)

- Wide spatial distribution of TCCON sites, can be tailored to improve coverage in sparse areas

- Network is relatively new but builds upon existing TCCON retrieval and validation methodology

Quality

This network has been shown to have minimal bias and long-term drift relative to TCCON. Uncertainties (2σ) of 0.6 µmol·mol^{-1} for CO_2 and 2.2 nmol·mol^{-1} for CH_4 stated but may vary depending on site.

A.2. In situ Methods

Description

Although less common than ground-based validation, direct validation of atmospheric satellite data within situ observations has been carried out previously. For example, in situ observations of CO_2 mole fraction from aircraft profiles have been directly compared with GOSAT and OCO-2 total CO_2 column (following extrapolation of aircraft profile data to top of atmosphere with model data). Good agreement was found between extrapolated aircraft CO_2 profiles and satellite retrieved CO_2 columns (Mustafa et al., 2021). In situ observations from dropsonde probes and balloon-borne sondes have also been used to validate atmospheric satellite data products (Baier et al., 2023; Mustafa et al., 2021).

In some cases, particularly with aircraft in situ observations, the uncertainties in the validation dataset are much lower than with remotely sensed atmospheric data (i.e. TCCON/COCCON or satellites). In situ validation also provides a more direct traceability link to established in situ calibration scales (e.g., WMO) than ground-based remote sensing methods. However, in situ validation activities are often sporadic and carried out in a case study-like fashion rather than as part of a formalized network. Such validation efforts therefore often lack the spatial and temporal coverage to be an effective validation strategy on their own. In situ observations are best utilized as supplementary validation datasets in support of more mature, widespread ground-based validation networks.

Scope of Representativeness

Considerations regarding the scope of representativeness include topics such as: There may be some degree of mismatch between aircraft/sonde profile and satellite column, as validation data must be extrapolated to match full vertical atmospheric column. There may be limited spatial and temporal coverage, as studies often performed on a case-by-case basis.

Quality

Data quality expectations are variable, depending on in situ technique, specific study, etc. Example 1σ in situ instrument precisions: 0.02 $\mu mol \cdot mol^{-1}$ for CO_2, 0.5 $nmol \cdot mol^{-1}$ for CH_4 (Wunch et al., 2010).

Appendix B. NASA Data Maturity Levels

Note that the following is also available at:

https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-and-information-policy/data-maturity-levels

Beta

Products intended to enable users to gain familiarity with the parameters and the data formats.

Provisional

Product was defined to facilitate data exploration and process studies that do not require rigorous validation. These data are partially validated, and improvements are continuing; quality may not be optimal since validation and quality assurance are ongoing.

Validated

Products are high-quality data that have been fully validated and quality checked, and that are deemed suitable for systematic studies such as climate change, as well as for shorter term, process studies. These are publication quality data with well-defined uncertainties, but they are also subject to continuing validation, quality assurance, and further improvements in subsequent versions. Users are expected to be familiar with quality summaries of all data before publication of results; when in doubt, contact the appropriate instrument team.

- **Stage 1 Validation:** Product accuracy is estimated using a small number of independent measurements obtained from selected locations and time periods and ground-truth/field program efforts.

- **Stage 2 Validation:** Product accuracy is estimated over a significant set of locations and time periods by comparison with reference in situ or other suitable reference data. Spatial and temporal consistency of the product and with similar products has been evaluated over globally representative locations and time periods. Results are published in the peer-reviewed literature.

- **Stage 3 Validation:** Product accuracy has been assessed. Uncertainties in the product and its associated structure are well quantified from comparison with reference in situ or other suitable reference data. Uncertainties are characterized in a statistically robust way over multiple locations and time periods representing global conditions. Spatial and temporal consistency of the product and with similar products has been evaluated over globally representative locations and periods. Results are published in the peer-reviewed literature.

- **Stage 4 Validation:** Validation results for stage 3 are systematically updated when new product versions are released and as the time-series expands.

Appendix C. Acronyms & Abbreviations

APA	Absolute Positional Accuracy
ARD	Analysis Ready Data
ATBD	Algorithm Theoretical Basis Document
BBR	Band-to-Band Registration
BIPM	Bureau International des Poids et Mesures
BIRA-IASB	The Royal Belgian Institute for Space Aeronomy
CEOS	Committee on Earth Observation Satellites
COCCON	Collaborative Carbon Column Observing Network
CF	Climate & Forecast (Metadata Convention)
ECV	Essential Climate Variable
EDAP	Earthnet Data Assessment Pilot
EMIT	Earth Surface Mineral Dust Source Investigation
EO	Earth Observation
ESF	Edge Spread Function
ESA	European Space Agency
FRM	Fiducial Reference Measurement
FRM4GHG	Fiducial Reference Measurements for Ground-Based FTIR Greenhouse Gas Observations
FTIR	Fourier Transform InfraRed spectroscopy
FWHM	Full Width Half Maximum
GCP	Ground Control Point
GSFC	Goddard Space Flight Center
GUM	Guide to the Expression of Uncertainty in Measurements
GSD	Ground Sampling Distance
IMAP-DOAS	Iterative Maximum a Posteriori - Differential Optical Absorption Spectroscopy
IMEO	International Methane Emission Observatory (of UNEP)
JPL	Jet Propulsion Laboratory
L0	Level 0
L1	Level 1
L2	Level 2
L4	Level 4
LSF	Line Spread Function
MTF	Modulation Transfer Function
NASA	National Aeronautics and Space Administration

NetCDF	Network Common Data Format
NPL	National Physical Laboratory, UK
POD	Probability of Detection
PSF	Point Spread Function
PUG/PUM	Product User Guide/Manual
QA4ECV	Quality Assurance Framework for Essential Climate Variables
QA4EO	Quality Assurance Framework for Earth Observation
RER	Relative Edge Response
SAR	Synthetic Aperture Radar
SI	Système International (International System of Units)
SNR	Signal-to-Noise Ratio
SRF	Spectral Response Function
SSAI	Science Systems and Applications, Inc
SSR	Sensor Spatial Response
TCCON	Total Carbon Column Observing Network
TROPOMI	Tropospheric Monitoring Instrument
VIM	International Vocabulary of Metrology
VVUQ	Validation, Verification, Uncertainty Quantification
WMO	World Meteorological Organization

Appendix D. Definitions and Terms

Calibrated Radiance

The calibrated radiance serves as the starting point for many teams. Several important characteristics should be recorded alongside the radiance, including the spectral grid and details about spectral sampling, such as the instrument line shape (ILS) or the full width at half maximum (FWHM) of the spectral response function. Additionally, it is necessary to record the signal-to-noise ratio (SNR) of the radiance, or a measure of noise as a function of wavelength. Lastly, information about the instrument's spatial response is required.

Concentration Enhancement

In practice, concentration enhancement refers to the analysis step where the background concentration or enhancement field is defined, and pixels with concentrations elevated above this background are identified.

Detection Rate

In the context of plume detection, detection rate refers to the ratio of true positive detections to number of observations.

Estimated detection threshold

In Jacob et al. (2016), the term "detection limit" is used to refer to the lower limit of the measuring interval for emission quantification. To align with the language of the metrology community, we will refer to this as the estimated detection threshold. The measuring interval is defined in the JCGM as "set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental measurement uncertainty, under defined conditions". Prior to having measurement data that can be used to determine the lower limit of the measuring interval and probability of detection curve as per the definitions, an approach to estimate the lower limit of the measuring interval or the estimated detection threshold can be constructed.

False Positives

The term "false positive" refers to the detection of plumes that do not exist.

Ground Sampling Distance and Point Spread Function

GSD is defined as the distance between the centers of two adjacent samples or pixels on the ground while the point spread function (PSF), an array of information about the spatial distribution of light that is seen by each pixel, is an integrated description of the performance of the full imaging system. These characteristics vary with parameters such as orbit altitude, the off-nadir angle, cross track and along track instrument characteristics, and other factors.

The GSD and the PSF are generally characterized pre-launch in the laboratory. This information must be combined with the location of the sensor (e.g., the position along an orbit track of a satellite) and its viewing direction to geo-locate each footprint relative to a surface map. In-flight verification can be performed using ground features such as coastlines, bridges, or small, isolated landmarks to assess the spatial sampling performance. Understanding of GSD and PSF

is crucial when determining the location of emission sources, as it significantly contributes to the uncertainty in source location. Source location data should always include the GSD and/or PSF characteristics, as it directly influences the precision of source geolocation.

Probability of Detection (POD)
POD is the term used in this community for capturing information about a measurement system's ability to detect methane plumes of various emission rates. The formal definition corresponds to the definition of detection limit provided by the Joint Committee for Guides in Metrology (JCGM). This definition relies on knowledge of the probabilities of false positives and false negatives. The POD is not a single value, but a set of values as a function of key driving variables.

JCGM Definition of Detection Limit which corresponds to our term Probability of Detection
The probability of detection is defined as the measured quantity value obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component is β, given a probability α of falsely claiming its presence.

JCGM Note 1: IUPAC recommends default values of 0.05 for both β and α.

JCGM Note 2: The term 'LOD' (limit of detection) is sometimes used.

JCGM Note 3: The term 'sensitivity' is discouraged when referring to detection limits.

In the application to methane plume detection, term β refers to the false negative rate, and the term α refers to the false positive rate. Note that the determination of the POD curve does not require quantification of plume emissions by the measurement system. If controlled release data us used, the POD curve can be determined from the controlled release known emission rates and the detects and non-detects of the measurement system.

Radiance Uncertainty
Radiance uncertainty arises from random and systematic sources of error. The radiance precision (random error) is driven by several factors, which are determined by the instrument's characteristics, including detector noise, detector efficiency, transmission efficiency, signal level, and integration time. Systematic errors may be related to offsets in electronics, error in calibration, systematic changes in the instrument, unaccounted for temperature sensitivity, and other factors.

Retrieval Uncertainty
Retrieval uncertainty is critical because, during the process of identifying enhancements relative to the background, both the uncertainty and the concentration resolution granularity will influence the results. As with radiance uncertainty, it arises from random and systematic sources of error. Two working definitions of retrieval random error (precision) are proposed. No details on the characterization of systematic error have been captured in this report.

Bayesian Retrieval Precision
This approach uses the posterior error covariance from an optimal estimation retrieval.

Empirical (Background) Retrieval Precision

This precision is empirically estimated based on column retrievals, obtained through replicate measurements on the same or similar objects under specified conditions, typically using background measurements where no methane plumes are present.

6. References

[1] Baier, B. C., Sweeney, C., & Chen, H. (2023). Chapter 8—The AirCore atmospheric sampling system. In N. R. Nalli (Ed.), Field Measurements for Passive Environmental Remote Sensing (pp. 139–156). Elsevier. https://doi.org/10.1016/B978-0-12-823953-7.00014-9.

[2] Bell, C., & Zimmerle, D. (2020). METEC controlled test protocol: continuous monitoring emission detection and quantification.

[3] Bouvet, M., et al. (2019), RadCalNet: A Radiometric Calibration Network for Earth Observing Imagers Operating in the Visible to Shortwave Infrared Spectral Range, *Remote Sens.*, 11(20), 2401, doi:10.3390/rs11202401.

[4] Burgdorf, M., Hans, I., Prange, M., Mittaz, J., and Woolliams, E., "FIDUCEO D2.2 (Microwave): Report on the MW FCDR Uncertainty," 2019. [Online]. Available: https://cordis.europa.eu/project/id/638822/results.

[5] CEOS ARD (2021), Minimum Product Family Specifications, Version 1, Available at: https://ceos.org/ard/files/PFS/CEOS-ARD_PFS_Template.docx.

[6] CEOS LSI, "CARD4L Product Family Specification - Surface Reflectance," 2020. [Online]. Available: https://ceos.org/ard/files/PFS/SR/v5.0/CARD4L_Product_Family_Specification_Surface_Reflectance-v5.0.pdf.

[7] Cusworth, D. H., Jacob, D. J., Varon, D. J., Chan Miller, C., Liu, X., Chance, K., et al. (2019). Potential of next-generation imaging spectrometers to detect and quantify methane point sources from space. Atmospheric Measurement Techniques, 12(10), 5655–5668. https://doi.org/10.5194/amt-12-5655-2019

[8] Cusworth, D. H., Thorpe, A. K., Ayasse, A. K., Stepp, D., Heckler, J., Asner, G. P., et al. (2022). Strong methane point sources contribute a disproportionate fraction of total emissions across multiple basins in the United States. *Proceedings of the National Academy of Sciences*, *119*(38), e2202338119. https://doi.org/10.1073/pnas.2202338119

[9] Darynova, Z., Blanco, B., Juery, C., Donnat, L., & Duclaux, O. (2023), Data assimilation method for quantifying controlled methane releases using a drone and ground-sensors, *Atmospheric Environment: X*, 17, 100210.

[10] Datla, R. U., Rice, J. P., Lykke, K. R., Johnson, B. C., Butler, J. J., and Xiong, X. (2011), Best practice guidelines for pre-launch characterization and calibration of instruments for passive optical remote sensing, *J. Res. Natl. Inst. Stand. Technol.*, 116(2), 621, doi:10.6028/jres.116.009.

[11] Dechoz, C., et al. (2015), Sentinel 2 global reference image, in Bruzzone, L. (ed.), *Image and Signal Processing for Remote Sensing XXI*, pp. 94–107, doi:10.1117/12.2195046.

[12] Duren, R. M., Thorpe, A. K., Foster, K. T., Rafiq, T., Hopkins, F. M., Yadav, V., et al. (2019). California's methane super-emitters. Nature, 575(7781), 180–184. https://doi.org/10.1038/s41586-019-1720-3

[13] Eaton, B., et al. (2020), NetCDF Climate and Forecast (CF) Metadata Conventions, Available at: https://cfconventions.org/latest.html.

[14] El Abbadi, S. H., et al. (2024), Technological Maturity of Aircraft-Based Methane Sensing for Greenhouse Gas Mitigation, *Environ. Sci. Technol.*, 58, 9591–9600.

NIST IR 8575
May 2025
er>

[15] Fox, N. (2019), FRM4STS D-180 Final Report, Available at: http://www.frm4sts.org/wp-content/uploads/sites/3/2020/01/OFE-D-180-V1-Iss-1-Ver-1-signed.pdf.

[16] Frankenberg, C., Thorpe, A. K., Thompson, D. R., et al. (2016), Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region, *Proc. Natl. Acad. Sci.*, 113(35), 9734–9739, doi:10.1073/pnas.1605617113.

[17] Frey, M., et al. (2019), Building the Collaborative Carbon Column Observing Network (COCCON): Long-term stability and ensemble performance of the EM27/SUN Fourier transform spectrometer, *Atmos. Meas. Tech.*, 12, 1513-1530, doi:10.5194/amt-12-1513-2019.

[18] Gorroño, J., et al. (2017), A radiometric uncertainty tool for the Sentinel-2 mission, *Remote Sens.*, 9(2), 178, doi:10.3390/rs9020178.

[19] Gorroño, J., Varon, D. J., Irakulis-Loitxate, I., and Guanter, L. (2023), Understanding the potential of Sentinel-2 for monitoring methane point emissions, *Atmos. Meas. Tech.*, 16, 89–107, doi:10.5194/amt-16-89-2023.

[20] Gutman, G., et al. (2013), Assessment of the NASA–USGS Global Land Survey (GLS) datasets, *Remote Sens. Environ.*, 134, 249–265, doi:10.1016/j.rse.2013.02.026.

[21] Hunt, S. E. (2021), Earth Observation Mission Quality Assessment Framework.

[22] INSPIRE Drafting Team Metadata and European Commission Joint Research Centre (2013), INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, Available at: https://inspire.ec.europa.eu/documents/inspire-metadata-implementing-rules-technical-guidelines-based-en-iso-19115-and-en-iso-1.

[23] Jacob, D. J., Turner, A. J., Maasakkers, J. D., Sheng, J., Sun, K., Liu, X., et al. (2016). Satellite observations of atmospheric methane and their value for quantifying methane emissions. Atmospheric Chemistry and Physics, 16(22), 14371–14396. https://doi.org/10.5194/acp-16-14371-2016

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

[25] JCGM (2008), Evaluation of measurement data - Guide to the expression of uncertainty in measurement, JCGM 100, Available at: https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

[26] JGCM (2012), International vocabulary of metrology – Basic and general concepts and associated terms (VIM), JGCM 200.

[27] Johnson, M. R., Tyner, D. R., & Szekeres, A. J. (2021), Blinded evaluation of airborne methane source detection using Bridger Photonics LiDAR, *Remote Sens. Environ.*, 259, 112418.

[28] Jongaramrungruang, S., Frankenberg, C., Matheou, G., Thorpe, A. K., Thompson, D. R., Kuai, L., and Duren, R. M. (2019), Towards accurate methane point-source quantification from high-resolution 2-D plume imagery, *Atmos. Meas. Tech.*, 12, 6667–6681, doi:10.5194/amt-12-6667-2019.

[29] Jongaramrungruang, S., Matheou, G., Thorpe, A. K., Zeng, Z. C., & Frankenberg, C. (2021). Remote sensing of methane plumes: instrument tradeoff analysis for detecting and quantifying local sources at global scale. Atmospheric Measurement Techniques Discussions, 2021, 1-29. doi:10.5194/amt-14-7999-2021

er>80er>

[30] Kunkel, W. M., Carre-Burritt, A. E., Aivazian, G. S., Snow, N. C., Harris, J. T., Mueller, T. S., … & Thorpe, M. J. (2023). Extension of methane emission rate distribution for Permian Basin oil and gas production infrastructure by aerial LiDAR. *Environmental Science & Technology*, *57*(33), 12234-12241. https://pubs.acs.org/doi/10.1021/acs.est.3c00229.

[31] Messerschmidt, J., et al. (2011), Calibration of TCCON column-averaged CO2: The first aircraft campaign over European TCCON sites, *Atmos. Chem. Phys.*, 11, 10765–10777, doi:10.5194/acp-11-10765-2011.

[32] Mittaz, J., Merchant, C. J., and Woolliams, E. R. (2019), Applying principles of metrology to historical Earth observations from satellites, *Metrologia*, 56(3), 032002, doi:10.1088/1681-7575/ab1705.

[33] Mustafa, F., et al. (2021), Validation of GOSAT and OCO-2 against in situ aircraft measurements and comparison with CarbonTracker and GEOS-Chem over Qinhuangdao, China, *Remote Sens.*, 13, 1–15, doi:10.3390/ rs13050899.

[34] National Imagery and Mapping Agency (2000), Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems, 3rd edn., Available at: http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html.

[35] Nguyen, H., Cressie, N., & Hobbs, J. (2019), Sensitivity of optimal estimation satellite retrievals to misspecification of the prior mean and covariance, with application to OCO-2 retrievals, *Remote Sens.*, 11, 2770, doi:10.3390/rs11232770.

[36] Nightingale, J., et al. (2019), Ten Priority Science Gaps in Assessing Climate Data Record Quality, *Remote Sens.*, 11(8), 986, doi:10.3390/rs11080986.

[37] Poursanidis, K., Sharanik, J., & Hadjistassou, C. (2024). World's largest natural gas leak from nord stream pipeline estimated at 478,000 tonnes. iScience, 27(1), 108772. https://doi.org/10.1016/j.isci.2023.108772

[38] QA4EO Task Team (2010), Quality Assurance for Earth Observation Principles, Available at: http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf.

[39] Redout-Leduc, G., Jacob, D. J., Varon, D. J., et al. (2024), Automated detection of methane point source plumes using deep learning applied to satellite imagery, *Atmos. Meas. Tech.*, 17, 765–782, doi:10.5194/amt-17-765-2024.

[40] Scanlon, T. (2017), QA4ECV Product Documentation Guidance: Provenance Traceability Chains, Available at: http://www.qa4ecv.eu/sites/default/files/QA4ECV%20Traceability%20Chains%20Guidance.pdf.

[41] Sherwin, E. D., et al. (2023), Single-blind validation of space-based point-source detection and quantification of onshore methane emissions, *Sci. Rep.*, 13, 3836.

[42] Sherwin, E. D., et al. (2024a), Single-blind test of nine methane-sensing satellite systems from three continents, *Atmos. Meas. Tech.*, 17, 765–782.

[43] Sherwin, E. D., Rutherford, J. S., Zhang, Z., Chen, Y., Wetherley, E. B., Yakovlev, P. V., et al. (2024b). US oil and gas system emissions from nearly one million aerial site measurements. *Nature*, *627*(8003), 328–334. https://doi.org/10.1038/s41586-024-07117-5

[44] Storey, J., Choate, M., and Lee, K. (2014), Landsat 8 Operational Land Imager On-Orbit Geometric Calibration and Performance, *Remote Sens.*, 6(11), 11127–11152, doi:10.3390/rs61111127.

[45] Thorpe, A. K., Frankenberg, C., Thompson, D. R., Duren, R. M., Aubrey, A. D., Bue, B. D., ... & Dennison, P. E. (2017). Airborne DOAS retrievals of methane, carbon dioxide, and water vapor concentrations at high spatial resolution: application to AVIRIS-NG. *Atmospheric Measurement Techniques*, *10*(10), 3833-3850.

[46] Thorpe, A., Green, R., Thompson, D., Brodrick, P., Chapman, J., Elder, C., et al. (2023), Mapping methane and carbon dioxide point sources from space with EMIT, *Science Advances*, doi:10.5194/egusphere-egu23-9429.

[47] United Nations Environment Programme (UNEP) (2021), International Methane Emissions Observatory, Available at: https://www.unep.org/topics/energy/methane/international-methane-emissions-observatory.

[48] Vendt, R. (2020), FRM4SOC D-290 Final Report, Available at: https://frm4soc.org/wp-content/uploads/filebase/parentdir/techreports/temp_pic/D-290-FRM4SOC-FR_30.06.2020.pdf.

[49] Wilkinson, M. D., et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 3(1), doi:10.1038/sdata.2016.18.

[50] Wolfe, R. E., et al. (2013), Suomi NPP VIIRS prelaunch and on-orbit geometric calibration and characterization, *J. Geophys. Res. Atmos.*, 118(20), 11,508–11,521, doi:10.1002/jgrd.50873.

[51] Wunch, D., et al. (2011), The total carbon column observing network, *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 369, 2087–2112.

[52] Wunch, D., et al. (2010), Calibration of the total carbon column observing network using aircraft profile data, *Atmos. Meas. Tech.*, 3, 1351–1362.

[53] Wunch, D., et al. (2017), Comparisons of the Orbiting Carbon Observatory-2 (OCO-2) XCO2 measurements with TCCON, *Atmos. Meas. Tech.*, 10, 2209–2238.

[54] Zimmerle, D., Dileep, S., & Quinn, C. (2024). Unaddressed Uncertainties When Scaling Regional Aircraft Emission Surveys to Basin Emission Estimates. Environmental Science & Technology, 58(15), 6575–6585. https://doi.org/10.1021/acs.est.3c08972