1.0 Author Page

Commented [PG1]: Need to make a suitable title page.

Starting Title: Common Practices for Quantifying, Reporting, Validating, and Assessing Facility Scale Methane Emissions Using Remote Sensing.

Lead Authors: John Worden[1], Paul Green[2], Annmarie Eldering[3], Evan Sherwin[4], and Adam Brandt[5]

1) NASA Jet Propulsion Laboratory / California Institute for Technology (USA)
2) National Physical Laboratory (United Kingdom)
3) National Institute for Standards and Technology (USA)
4) Lawrence Berkeley National Laboratory (USA)
5) Stanford University (USA)

Contributing authors (section 3)

Jason McKeever, GHGSat
Dylan Jervis, GHGSat
Daniel Varon, Harvard
Andrew Thorpe, JPL, EMIT team
Philip Brodrick, JPL, EMIT team
James Whetstone, NIST
Riley Duren, Carbon Mapper

Contributing authors (section 4)

By Evan D. Sherwin[1], Adam R. Brandt[2]
[1] Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area and Energy Technologies Area
[1] Stanford University, Department of Energy Science & Engineering

CONTRIBUTING AUTHORS (section 5)
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

Executive Summary:

This document is the outcome of an initial activity to collate and articulate best practice for generating, validating, and assessing the quality of facility-scale methane emission identification & estimates derived from spectroscopic remote sensing radiances[1]. In this initial phase it is a collation of existing practices, distilled to identify common practices & considerations. These are presented as generalized community-prevalent practices and considerations best suited to a robust identification and quantification of facility-scale methane emission.

Commented [2]: Limiting to spectroscopic - presume the widest definition, where GOES/LandSat are still 'multi-channel' spectrometers or are we only thinking hyperspectral and high-spectral resolution dedicated missions?

Commented [3]: Spectroscopic is very general term.. here intended to mean that you are looking at a radiance that is modified by a methane band. I think that covers what we are talking about.

This document is intended for both producers and users of remotely sensed facility-scale methane emission identification & quantification products. Producers can use this document as a guide to understand the standards expected by practitioners in the field before their data can be considered reliable. Users can refer to this document to identify key elements of the data, including considerations that could guide their expectations and allow a balanced assessment as to whether a dataset is fit-for-purpose to their need.

The document is organized as follows:

- **Section 2** describes the motivation and timeliness to develop & articulate best practice in remotely-sensed facility-scale methane emission identification & estimates
- **Section 3 describes** the current, community-common practices for quantifying methane emissions based on the measured radiances.
- **Section 4** outlines the current state-of-practice for validating these observations.
- **Sections 5 and 6** provide a template for quality assessment of the column methane values and emissions respectively; these can be used to evaluate facility scale methane emissions estimates according to expectations from the satellite community.

2.0 Background

In response to international agreements, including the COP26 Methane Pledge (Global Methane Pledge, 2021) and the COP28 Oil and Gas Decarbonization Charter (COP28 UAE, 2023), the reduction of fugitive methane emissions from industrial

[1] For clarity, remotely sensed spectroscopic radiances are considered in the most general sense, as the use of multiple spectrally defined radiances (coincident or not to methane sensitive features in this case) used in combination to derive some knowledge of emitted methane.

processes is high on the international agenda. Over the last few years, several initiatives have emerged, such as the launch of the UNEP International Methane Emissions Observatory (IMEO) program (UNEP, 2021) and the World Meteorological Organization (WMO) Global Greenhouse Gas Watch (G3W) initiative (WMO, 2022), along with the development and deployment of new methane monitoring capabilities (Jacob et al., 2022). These actions are being enshrined in national targets, policies, and, most recently, new regulations. Recent legislation in the United States (White House, 2022) and the European Union (IEA, 2023) not only mandates verifiable reductions in emissions but also promotes the use of innovative methods and techniques for identifying and reporting leaks, as well as performing third-party verification of the oil and gas sectors' emission reporting to regulators. Information is also being shared with federal and state agencies (such as the US EPA and the State of California), as well as through platforms directed at the companies and businesses responsible for emissions.

Simultaneously, publicly listed companies are being required to report their emissions, as well as their physical and financial risks related to climate resilience and operating in a low-carbon economy. Although more derived in its application, facility- and asset-level emissions data are now part of regulatory reporting requirements, with significant financial penalties for non-compliance or gross inaccuracies.

These diverse goals can only be quantifiably met with verified data, where satellite-derived measurements play a key role. The global reach and inherent spatial sampling and mapping capabilities of on-orbit instruments make them ideal for conducting consistent surveys across borders, cataloging sources (and sinks) of greenhouse gas emissions, and pinpointing their geographic locations.

To address this data need, a number of new on-orbit sensors from commercial and philanthropic (new space) stakeholders are joining longer-running public missions that track methane concentrations at a range of spatial resolutions (from tens of meters to a few kilometers). Innovations have also demonstrated that some public missions not originally designed to monitor methane can still do so, though they are limited to detecting more intense point sources. This expansion of satellite data has allowed multiple players to enter the methane emissions product landscape, including start-ups, academia, space agencies, on-orbit asset owners, and international organizations. To improve transparency and usability, both public and privately generated data are now being curated and integrated by national agencies such as Copernicus, the USA GHG Center, and the Japanese GHG Center.

While this dynamic influx of new data is welcome, it is also vulnerable to several challenges. In a relatively immature and rapidly developing field, divergent emissions estimates from various actors could undermine credibility. Additionally, questionable methods and data quality from non-expert players could significantly damage the reputation of the entire sector.

Community-accepted best practices would provide a baseline for comparison and support the adoption of these new data. Compliance with best practices would endorse reputable suppliers, filter out bad actors, and give the necessary confidence to the user and customer base.

Internationally adopted standards based on transparency, traceability, independence, and evidence-based quality assurance (QA) would ensure that the data is fit for purpose and demonstrate interoperability between suppliers.

To consolidate and articulate best practices for emissions quantification, reporting, and validation, the greenhouse gas (GHG) community, through the Committee on Earth Observation Satellites (CEOS) and National Metrology Institutes (NMIs), has identified the need for a "Best Practices" document. This document would outline community-common practices from L0/L1 (radiance) to L2 (concentration) to L4 (emissions) and include the current state-of-practice for validating these measurements.

The advent of new space missions (non-public missions) and the increasing use of their products by public entities also necessitate a "quality" assessment of these products. New space measurements and associated proprietary methods often create barriers to transparency, limiting the disclosure of the full data chain from L0 to L4 or the corresponding algorithms.

For this reason, CEOS and the NMIs have initiated a "Best Practices" effort, initially focusing on the measurements of facility-scale methane concentration plumes and corresponding emissions (spatial scales of approximately 10 meters). The Best Practices document will include community-accepted algorithms for L0 to L4, the state-of-the-art for validating facility-scale emissions estimates, and a template for assessing the quality of reported emissions products.

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

3.1 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 *super-emitter* emissions. One key motivation for consolidating these practices is the significant variability in emission estimates by different groups working with the same radiance measurements. To fully understand the root causes of this variability, we aim to document and analyze the processes in detail. An important outcome of collecting these common practices is the establishment of a shared vocabulary. Thus, we document agreed-upon definitions and link them to reference materials, such as the Joint Committee for Guides in Metrology's *Guide to the Expression of Uncertainty in Measurement* (JCGM GUM), published by the BIPM.

All the instruments discussed here use hyperspectral imagers that gather measurements of reflected sunlight in the spectral regions where methane absorbs light (around 1.6 and 2.3 microns). 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 to 1 nm. Several papers document the range of instrumentation and missions, such as Jacob et al. (2022).

The radiance measurements are then used to estimate methane concentrations in the area of interest. This is generally done through a physics-based retrieval approach (e.g., IMAP-DOAS) or a statistical method, such as a matched filter (Thorpe et al. 2023). 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 (Ueff, m s−1), and a plume length scale (L, m).

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

For further discussion of the measurement concepts and additional background information, readers are encouraged to refer to the following resources:

- Carbon Mapper FAQ
- UNEP International Methane Emissions Observatory
- Kayrros Technology Overview
- NASA Methane Source Finder
- Bridger Photonics Methane Detection

3.2 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 sub-sections, each step is described in detail, including definitions and current practices. We also note some outstanding issues and unresolved questions throughout.

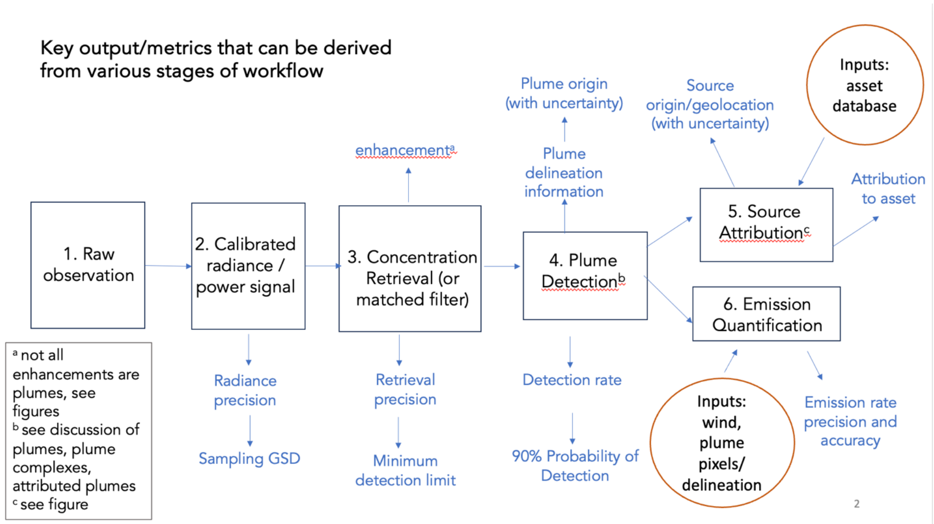

Key output/metrics that can be derived from various stages of workflow
1. Raw observation
2. Calibrated radiance / power signal
3. Concentration Retrieval (or matched filter)
4. Plume Detection[b]
5. Source Attribution[c]
6. Emission Quantification
enhancement[a]
Plume origin (with uncertainty)
Plume delineation information
Source origin/geolocation (with uncertainty)
Inputs: asset database
Attribution to asset
Radiance precision
Sampling GSD
Retrieval precision
Minimum detection limit
Detection rate
90% Probability of Detection
Inputs: wind, plume pixels/ delineation
Emission rate precision and accuracy
[a] not all enhancements are plumes, see figures
[b] see discussion of plumes, plume complexes, attributed plumes
[c] see figure
2

Figure 1. Typical analysis steps for methane plume detection and quantification process. Credit: Dan Cusworth, Carbon Mapper

3.3 Calibrated radiance/ power signal element

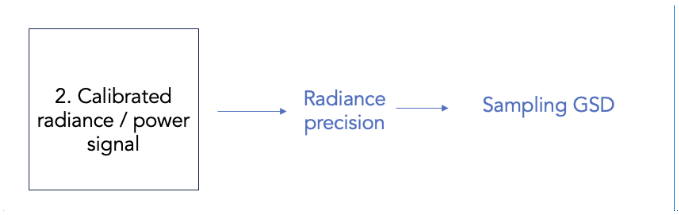


Figure 2 L1 data Credit: Dan Cusworth, Carbon Mapper

Commented [PG4]: @Annmarie We need to rename radiance precision to radiance uncertainty - precision is effectively repeat measurement consistency - in practise a measure of measurement noise. Apologies - didn't notice this first time through.

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

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.

In our generalized workflow figure, only radiance uncertainty and the ground sampling distance (GSD) are noted and discussed here. Additional characteristics will be included in future updates.

Radiance Uncertainty

Radiance uncertainty arises from several factors, which are determined by the instrument’s characteristics, including detector noise, detector efficiency, transmission efficiency, and integration time. 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 background concentration field variations. Typically, emphasis is placed on the linearity of the measurement system. Some in-flight radiometric verification can be conducted, using instrumented sites like those provided by RadCalNet.

Sampling GSD

Ground sampling distance (GSD), defined as the distance between the centers of two adjacent samples on the ground, varies with the off-nadir angle. Instrument teams typically report GSD for nadir views, often as the full width at half maximum (FWHM) of a Gaussian approximation for the spatial sampling response function.

Both the GSD and the spatial response function are generally characterized pre-launch in the laboratory. In-flight verification can be performed using ground features such as coastlines, bridges, or small, isolated landmarks to assess the spatial sampling performance. Understanding GSD 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, as it directly influences the precision of source geolocation.

3.4 Concentration retrieval/matched filter

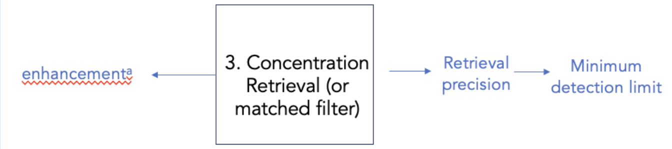


Figure 3. Concentration step contributors. Credit: Dan Cusworth, Carbon Mapper

Commented [PG5]: Need to edit this too - precision here is more correct, but is one aspect only.

This section provides an overview of current approaches used for concentration retrievals (Figure 3) in methane detection workflows. As the field rapidly evolves, several techniques have emerged, each with varying levels of accuracy and application. Additionally, we highlight the importance of retrieval uncertainty and common practices related to concentration enhancements.

Concentration Retrievals

There are several different approaches currently in use for concentration retrievals. As this field continues to evolve, some notable examples of techniques include:

- **Band Difference/Band Ratio** (Frankenberg et al., 2016)
- **Matched Filters** (Foote et al., 2017)
- **Full Physics** (Boesch et al., 2011)

Ideally, the concentration retrieval process also provides an estimate of retrieval uncertainty to ensure robust analysis. Figure 3 illustrates the concentration retrieval step within the broader methane detection workflow.

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. Generally, higher retrieval uncertainty limits the ability to identify smaller enhancements.

Two working definitions of retrieval precision are proposed, depending on whether optimal estimation (OE) retrieval is performed. These definitions are outlined below:

- **Bayesian Retrieval Precision** – This approach uses the posterior error covariance from an optimal estimation retrieval.
    - *Notes:*
        - In OE retrievals, constraints, priors, and other assumptions can significantly affect the retrieval uncertainty.
        - A discussion of the impact of priors and prior misspecification can be found in Nguyen et al. (2019).
        - Column precision 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 may underestimate the impact of artifacts and unmodeled physical effects.
        - 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.

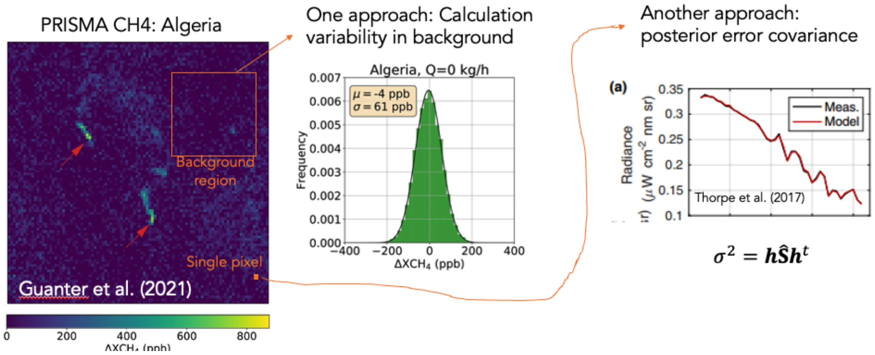


Figure 4 Concentration precision step. Credit: Dan Dan Cusworth, Carbon Mapper

- **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.
    - *Notes:*
        - 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.
        - When assessing methane enhancements above background levels, the mean value should be zero, and any variability reflects the uncertainty of the retrieved methane.
        - Observing conditions must also be documented. Ideally, empirical measurement precision should be calculated under the same albedo, solar zenith angle (SZA), and view angle, so the results are comparable. In the absence of this, any reported precision must include the viewing conditions.

Concentration Enhancement

In practice, concentration enhancement refers to the analysis step where the background concentration is defined, and pixels with concentrations elevated above this background are identified. 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 is a general concept that identifies pixels with elevated concentrations, which then feeds into the subsequent plume detection step, discussed later.

3.5 Plume detection

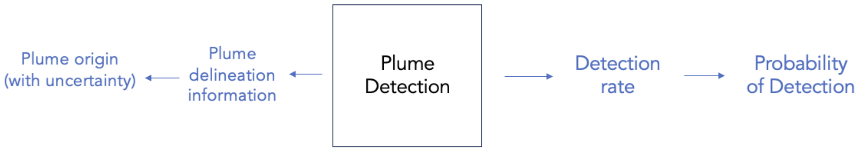

Plume origin (with uncertainty)
Plume delineation information
Plume Detection
Detection rate
Probability of Detection

Figure 5 plume detection step Credit: Dan Cusworth, Carbon Mapper

Plume segmentation or delineation is a critical step (Figures 5) in methane 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). 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.

Plume segmentation refers to the step where a set of enhanced pixels are selected and grouped to define the plume. During this step, additional parameters needed for plume quantification, such as the plume characteristic length (L), are developed.

Currently, there is no comprehensive review available that details the methods for plume segmentation or delineation. In many cases, the segmentation approach is presented alongside the quantification methods in scientific publications. However, this is an area of rapid technological advancement, with new techniques emerging regularly.

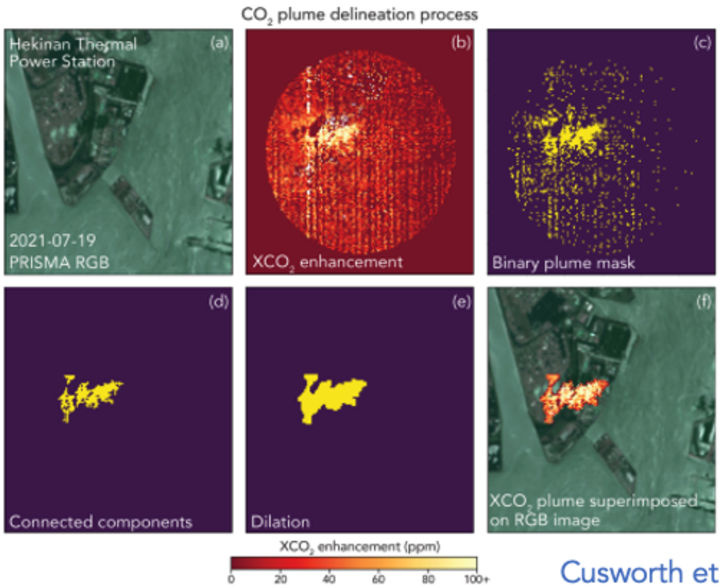
Common approaches being used include:

- 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, followed by quality control before the next steps. This manual review introduces bias, which could affect the probability of detection (POD). As a result, further development is needed in the best practices to mitigate this bias.

Key considerations for improving consistency and reducing operator bias include:

- Using a signal-to-error ratio
- Providing an error estimate for each pixel (empirical value)
- Accounting for scene artifacts, where possible
- Differentiating between artifacts and true enhancements


CO$_2$ plume delineation process
Hekinan Thermal Power Station
(a)
2021-07-19
PRISMA RGB
(b)
XCO$_2$ enhancement
(c)
Binary plume mask
(d)
Connected components
(e)
Dilation
(f)
XCO$_2$ plume superimposed on RGB image
XCO$_2$ enhancement (ppm)
0
20
40
60
80
100+

Cusworth et al. (2023)

Figure 6 Steps showing the process of finding the methane enhancement and then the plume.

Discussion of Current Practices

The consensus from the community is that a broad common approach is used for plume segmentation. 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 delineation for visualization purposes 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 delineation is practiced, typically with a high threshold to ensure accuracy.

Key Notes:

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

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

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

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

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

Plume Origin

Once the plume delineation 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.

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

- Concentration fields or matched filter outputs
- Wind direction
- The overall shape of the segmented plume (e.g., cone-shaped plumes)
- Surface imagery, including topographical features and infrastructure data

Key Observations:

- There is a significant variation in common practices for determining the plume origin.
- 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 paid 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.

3.6 Source attribution

Source attribution is a critical step (Figure 6) in methane emissions detection, as it links detected plumes to their emission sources. This section addresses the common practices and challenges faced in determining the origin of emissions and attributing them to specific assets. The accuracy of this step is key to ensuring that emissions are correctly assigned, whether for regulatory purposes, mitigation, or enforcement.

Source Origin/Geolocation

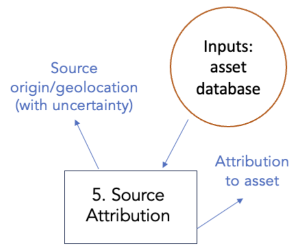


Figure 7 Source attribution Credit: Dan Cusworth, Carbon MapperSource Attribution: An Overview

Determining the source origin or geolocation of a plume can be complex (Figure 7), especially when multiple measurements are involved. In the case of a single observation, the plume origin is typically used as the source origin. However, when multiple observations are available, the mean location may be considered as the source origin.

For further details on how plume origins are developed, refer to the "Plume Origin" section.

Attribution to Asset

A distinction is generally 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.

However, 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 ground sampling distance (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.

3.7 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. Probability of detection provides a more accurate understanding of how likely a measurement system is to detect methane emissions under various conditions. This section explores the formal definitions of detection limits and POD, addresses challenges faced in the field, and outlines methods for evaluating these metrics. Additionally, it discusses the issue of false positives and the limitations of current detection systems.

Definitions

Detection Rate:
The detection rate offers some insight into the likelihood that a methane source will be detected by a measurement system. However, it is less precise than the preferred concept of probability of detection (POD).

Probability of Detection (POD):
POD is the preferred term for capturing information about a measurement system's ability to detect methane plumes of various emission rates. The formal definition provided by the Joint Committee for Guides in Metrology (JCGM) outlines the relationship between the probabilities of false positives and false negatives in this context.

JCGM Definition of Detection Limit:
The detection limit 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.

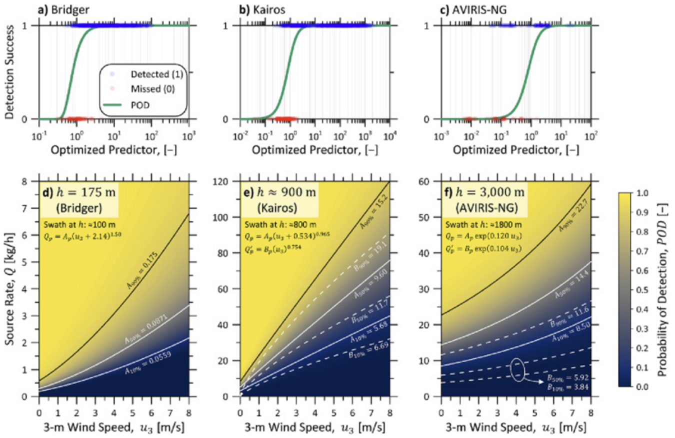
Challenges in Implementation

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

- The term "minimum detection limit" is often misinterpreted as the smallest emissions ever observed. For this reason, the community avoids using it.
- 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 F1 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.

To ensure comparability across different systems, teams should document observing conditions alongside their POD evaluations (Figure 8). Ideally, a standard set of reference conditions would be used across all teams.

Example of parametric POD curves derived from controlled release experiments



Conrad et al. (2023)

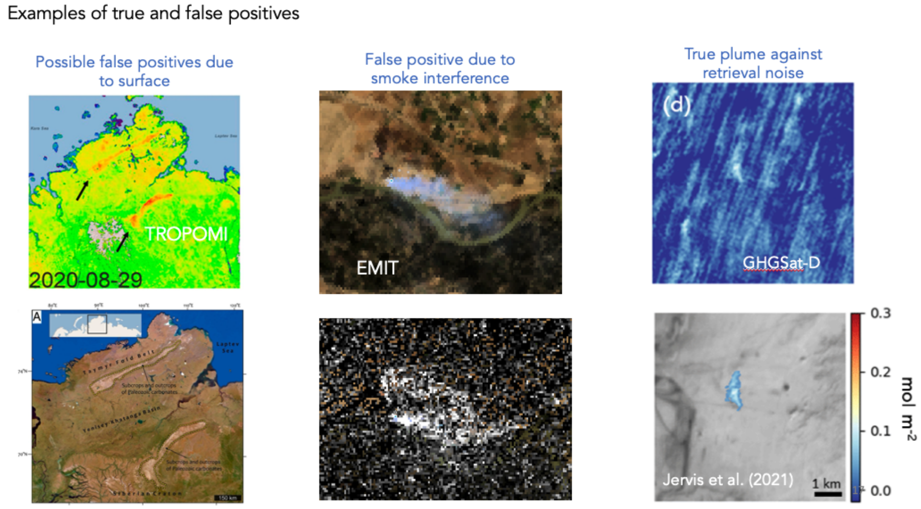
Figure 8

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).

False Positives:

False positives refers to the identification of plumes that do not truly exist. There are a number of reasons that this might occur, and some of those are illustrated below. For example, clouds or a smoke plume might be incorrectly identified as a methane plume. Surface features with surface reflectance that contrasts the background may be misidentified as a plume. 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, 2024).

Commented [6]: so current practise is a sniff test - no O&G infrastructure in the area, so a red herring. But the RGB image shows a land form/forest etc that can reasonably account. Just need to spell it out.


Examples of true and false positives

Possible false positives due to surface

False positive due to smoke interference

True plume against retrieval noise

Figure 6. False positive scenarios. Credit: Dan Cusworth, Carbon Mapper

Detection limit (in reference to emissions)

In this community, the language detection limit is used to refer to the lower limit of the measuring interval. 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 detection limit as per the definitions, an approach to estimate the lower limit of the measuring interval is as follows:

This is from the detection limit using mass balance arguments in (Jacob et al., 2016). The Jacob et. al. 2016 formula should be considered a "rule of thumb"/approximation that is particularly helpful when you don't have enough controlled release data to determine the POD curve empirically. This way you can tabulate *rough* DL values for many systems, planned and existing.

Q_{min} = (M_{CH4}U*W*Pa*q*r)/(g*M_a)

Q_{min} is the MDL (kg h^{-1}) [which we will refer to as detection limit]
W is the pixel size (in meters)
U is the wind speed (m s^{-1})
M_{CH4} is the molecular weight methane of (0.016 kg mol^{-1})
M_a is the molecular weight of air (0.029 kg mol^{-1}),
Pa is the dry atmosphere surface pressure ,
g is the acceleration due to gravity (9.8 m s^{-2}),
r 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). Tied to definition - We define detectability as a precision of delta-X/2 and quantification as a precision of delta-X/5. Delta-X is the mean enhancement

3.8 Emissions Quantification, ***Common Practices and Open Issues***

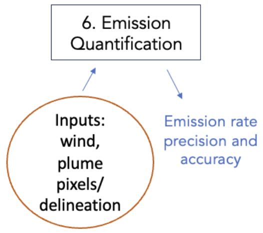

6. Emission Quantification
Inputs: wind, plume pixels/ delineation
Emission rate precision and accuracy

Figure 9. Emission quantification considerations. Credit: Dan Cusworth, Carbon Mapper

Emission rate quantification is a crucial aspect (Figure 9) of methane plume detection and mitigation. It 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.

Common Practices for Emission Rate Quantification (Including 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.

1. Integrated Mass Enhancement (IME):
   IME is used to calculate methane emissions by integrating the mass enhancement of methane in the detected plume. This method typically involves summing the concentration enhancements across the pixels of the plume and applying relevant corrections based on plume length and wind speed. The IME method is particularly useful for estimating emissions from large-scale plumes, but it requires accurate input data, especially regarding plume dimensions and wind conditions.
2. Cross-Sectional Flux (CSF):
   CSF focuses on estimating the flux of methane by calculating the cross-sectional methane concentration along a plume, combined with wind speed data. This method offers an alternative to IME and is especially useful when the spatial extent of the plume is well-defined, and wind data are reliable. However, the

accuracy of CSF can be affected by uncertainties in wind measurements and assumptions regarding plume structure.

Other quantification methods are also used in the community, 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.

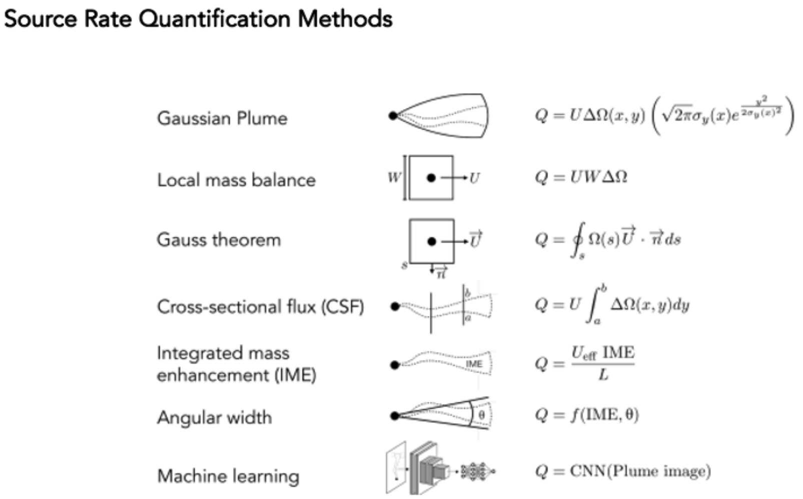

Source Rate Quantification Methods
Gaussian Plume
$Q = U\Delta\Omega(x,y)\left(\sqrt{2\pi}\sigma_y(x)e^{\frac{y^2}{2\sigma_y(x)^2}}\right)$
Local mass balance
W
U
$Q = UW\Delta\Omega$
Gauss theorem
$\vec{U}$
s
$\vec{n}$
$Q = \oint_s \Omega(s)\vec{U}\cdot\vec{n}\,ds$
Cross-sectional flux (CSF)
b
a
$Q = U\int_a^b \Delta\Omega(x,y)dy$
Integrated mass enhancement (IME)
IME
$Q = \frac{U_{\text{eff}}\ \text{IME}}{L}$
Angular width
θ
$Q = f(\text{IME}, \theta)$
Machine learning
$Q = \text{CNN(Plume image)}$

Figure 7. Emission estimate methods. From Jacobs et al 2022.

For clarity, including the equations and term definitions here, taken from Varon 2018.

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 – emissions estimate
Tau – residence time of methane in the detectable plume
IME – integrated mass enhancement
Ueff – 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 Ueff and L: L is a measure of plume extent, which can be interpreted as the length, perimeter of plume, or the square root of the area of the plume. Ueff is then derived from a simulation dataset, typically using large eddy simulations (LES). Given a source of wind data (U10 for example, where U10 is the 10 meter (altitude) product), and a choice of definition of L, an empirical relationship between Ueff and U10 is derived. The simulations are derived for specific instrument configurations (spatial resolution and noise), and also have specific atmospheric conditions, such as value or range of sensible heat flux and mechanical turbulence.

CSF equation:

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

Explanation of terms:
From Varon et al. (2018) paper: “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:
The integral is approximated in the observations as a discrete summation of the product U(x,y) * ΔΩ(x,y) over the detectable width of the plume.

But, a disadvantage 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 Ueff applied to the cross-plume integral C [kg m−1] of the column along the y axis.”

Current issues

The application of the IME equation requires L (plume length or plume area), Ueff, and IME, or the summed methane enhancement in the pixels included in the plume. L and Ueff are effectively co-dependant quantities, calibrated via LESs.

Common realization is that Ueff to U10 relationship is effectively a calibration, not a firm physics representation. The Ueff to U10 calibration curve also depends on the choice of definitions (L as a plume length or plume area), the source of windspeed (U10 may vary across data sources), plume delineation (which pixels are included) as well as the concentration retrieval (methane enhancement in the included pixels). Some discussion about how Ueff and L definitions that practitioners are using should be clearly reported, and they perhaps are not the best names for the terms, since Ueff is not really an effective wind, nor is L necessarily a measurable length.

A number of studies have employed LES simulations to derive Ueff to U10 relationships, starting with Varon 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. However, the conclusion from this set of work is that in the 1-5 m/s range the fits across papers/ensembles/noise levels only differ by 5%-15% or so". With the error incurred by not having a specifically derived relationship for each project is not that large (order 10%). (D Varon, private comms). However, very different results were observed for the work with the Nordstream data, with the differing conditions over ocean much of the plume obscured by cloud.

Ideally a large set of LES simulations covering all observing system, exploring a range of Ueff and U10 relationships, could sample the parameter space, but this is deemed impractical. The options currently considered are to a) perform simulations 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).

Comparing multiple methods provides some route to a sense check, comparing CSF and IMF quantifications, for instance, utilizing the advantages of both methods. CDF is not widely used in operational analysis, but is used in some cases for a 'sanity check' on the IME results.

One approach employed by one practitioner is to perform multiple retrievals on a range of plume lengths. After starting with the full plume (length), and performing IME, shorter plumes are analyses, (75%, 50%) and the variation in calculated Q used as a measure of uncertainty.

To better understand the uncertainties in the calculated Q, such permutation methods 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.

3.9 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. We propose the following framework for data product organization and contents.

1. Data Formats
    a. L1B – calibrated and geolocated radiance
        i. The preferred SI unit for radiance data is W/(m^2.µm.sr). Practitioners are also reporting in µW/(cm².sr.nm)
        ii. The wavelength or wavenumber grid must be included along with the radiance spectrum
        iii. Uncertainty on the radiance must also be included
        iv. information about the spectral response functions must be available, although it is not necessarily packaged with each radiance spectrum
        v. Similarly, information about the spatial response functions must be available, although it is not necessarily packaged with each radiance spectrum
    b. L2B – whole scene orthorectified atmospheric retrievals
        i. This is the output of the concentration retrieval step, which may be from a full physics retrieval approach or a match filter approach
        ii. The expected data units are ppm m^{-1} or kg per pixel
        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 identified as separate from the background.
    d. L3A and L3B - identified plumes
        i. These files include geotiffs 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/hour, 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

f. 4B – Source emission estimate, where they aggregate information over many observations. Would need significant ancillary information about what is included in the aggregation and the methodology

2. Data documentation

a. L1b radiance

i. Should be orthorectified, documented methods

ii. Radiance calibration procedure (if one exists) should be documented

iii. They should strive to relate their radiances to a known standard, using measurement sites such as RadCalNet

iv. Spectral calibration procedure should be described

b. L2B –

i. Document process for transforming radiance to total column or MML results

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

iv. Provide uncertainty on total column

v. If using MML, document all key parameters, in ATBD if fixed, per scene if they vary

c. L2C – detection

i. This is an area where there are a wide range of practices
Methodology should be described in ATBD or user guide

D. L3A and L3B -

This is an area where there are a wide range of practices
Methodology should be described in ATBD or user guide

4.0 The state of validation for point-source methane sensing satellite systems

4.1 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 the presence and quantity of methane emissions from a facility or location (L4 data).

Because of the focus on detection and quantification outcomes, the primary mechanism of validation for such systems has to date been single-blind controlled methane releases. We do not discuss traditional methods of satellite validation, such as Total Carbon Column Observing Network (TCCON) towers or aircraft data that provide a reference point for total column measurements at a particular point in space. These are important for understanding the performance of foundational data (e.g., L2) that are used as inputs for the methane emissions algorithm but are not applicable for tests of emissions detection and quantification outcomes.

4.2 Current Controlled Release Approach for Satellites

Existing satellite testing is of limited scope with first papers being published based on tests in 2021 and 2022. These tests have focused on a single-blind, known location design with emission rates ranging from 0.03 to 7.6 t(CH4)/h (Sherwin et al., 2023; Sherwin et al., 2024; 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 Sherwin et al. (2024), 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. 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 (Zimmerle and Bell, 2022).

Such testing provides insights into detection capabilities discussed in Section 3, including:

1. Providing an upper bound on the smallest emission the system is capable of detecting
2. Determining the presence/absence of false positive detections (reports of emissions when none were present)
3. 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, although this does not rule out the possibility of false positives in the field.

Point 3 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 100-150 kg per hour 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., 2024; Darynova et al., 2023).

4.3 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:

1. Characterizing any bias in quantification volumes across measurements

2. Characterizing uncertainty associated with a given measurement

4.4 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., 2024; 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 55-75% of measurements falling within ±50% of the metered value (Sherwin et al., 2023; Sherwin et al., 2024; El Abbadi et al., 2024; Bell et al., 2022).

In most methane remote sensing algorithms, the estimated emission rate is modeled as proportional to estimated wind speed, meaning 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., 2024). 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., 2024), 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., 2024).

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 1-16 days, past campaigns of 3-8 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.

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.

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)

Aggregated field statistics, such as those described in Kunkel et al. (2023), will likely contribute to estimating lower detection capabilities.

5.0 Introduction: 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.

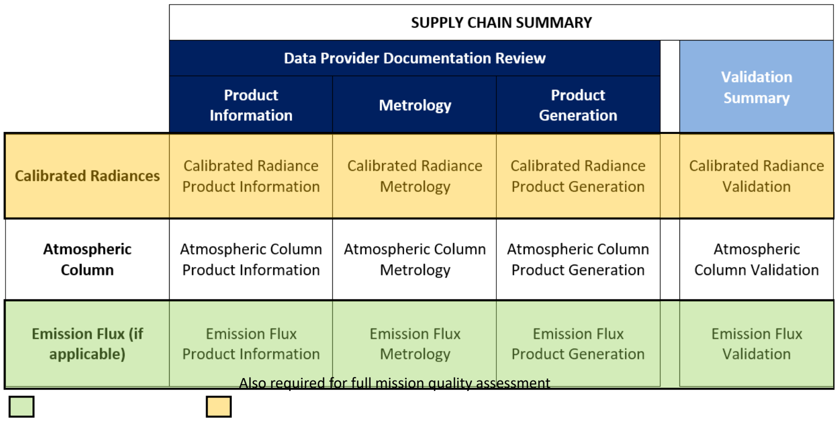
To ensure that decisions on commercial data acquisitions can be made fairly and with confidence, there is a need for an objective framework with which their data quality may be assessed. The ESA Earthnet Data Assessment Pilot (EDAP) project therefore defines this EO mission quality assessment framework 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 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 6) focuses on the methane emissions estimates made available from these measurements, in particular using the file containing plume enhancement values.

5.1 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:



	SUPPLY CHAIN SUMMARY			
	Data Provider Documentation Review			Validation Summary
	Product Information	Metrology	Product Generation	
Calibrated Radiances	Calibrated Radiance Product Information	Calibrated Radiance Metrology	Calibrated Radiance Product Generation	Calibrated Radiance Validation
Atmospheric Column	Atmospheric Column Product Information	Atmospheric Column Metrology	Atmospheric Column Product Generation	Atmospheric Column Validation
Emission Flux (if applicable)	Emission Flux Product Information	Emission Flux Metrology	Emission Flux Product Generation	Emission Flux Validation

Also required for full mission quality assessment

🟩 🟨

Figure 5-1: Supply Chain Summary for L1 (calibrated radiances), L2 (Atmospheric column) and L4 (Emission)

- *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 5.3 and 5.4, 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 summarised in a colour-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.

5.1.1 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 1). 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 1, 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 reproduceable from the reported L1B data and associated documentation (e.g. ATBDS and product user guides).

5.1.2 Quality Assessment Report

The quality assessment 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.

5.1.3 Product Evaluation Matrix

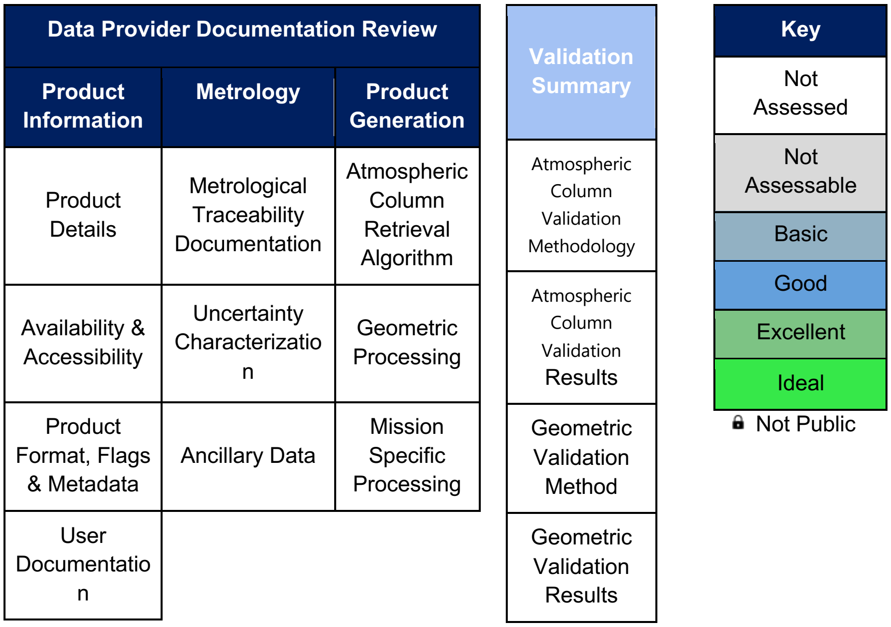
The product evaluation matrix provides a high-level colour-coded summary of the quality assessment results. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. Subsection grades are indicated by the colour of the respective grid cell, which are defined 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.

Summary Product Evaluation Matrix

The *Summary Product Evaluation Matrix* is shown in Figure 2. 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

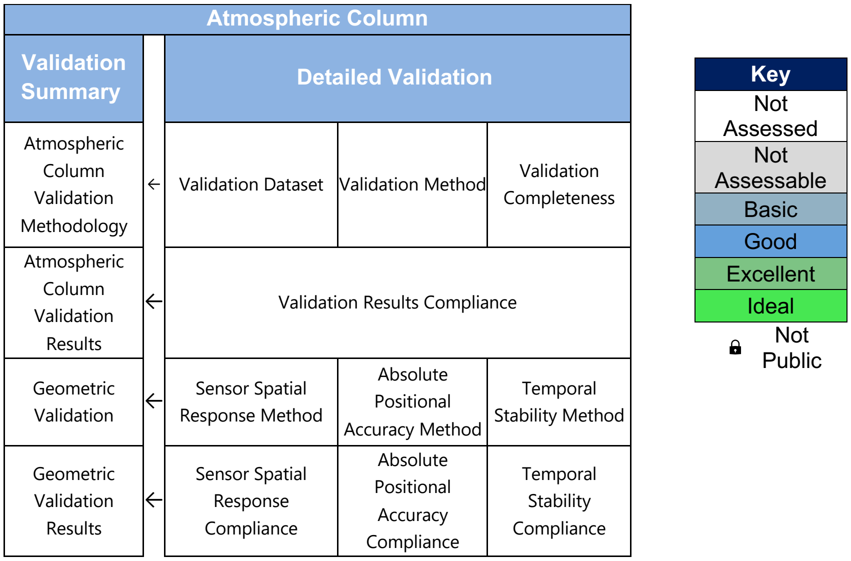
Key
Not Assessed
Not Assessable
Basic
Good
Excellent
Ideal

🔒 Not Public

Figure 5-2. Summary Product Evaluation Matrix.

Detailed Validation Maturity Matrix

The *Detailed Validation Maturity Matrix* (Figure 5-3) 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	←	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

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

Figure 5-3. Validation Maturity Matrix, showing the Validation Summary column from the Product Evaluation Matrix

5.2 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. The criteria for grading each box of the matrix are described in Sections 5.3 and 5.4

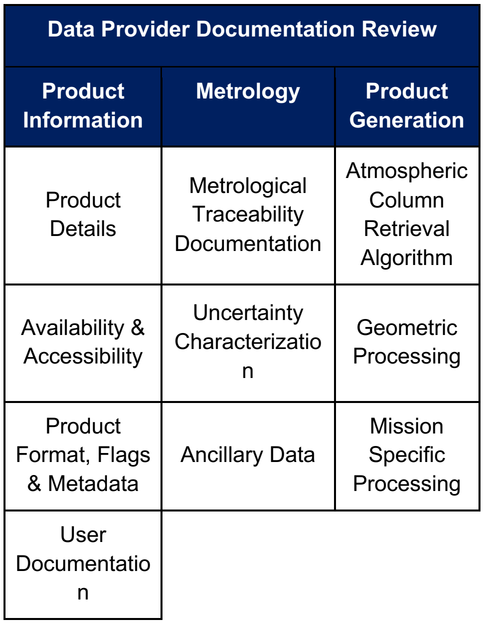
Additionally, a subsection may also indicate Not Assessable or Not Assessed. These cover the cases where certain aspects of product quality will not be assessed – either because there is insufficient information available to make an assessment, or because it is out of scope of the assessment.

5.3 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 5-2)*. This portion is shown in Figure 5-4.



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

Figure 5-4 – Data Provider Documentation Review Matrix

5.4 Product Information

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

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 organisation, including email address)
- Product access (e.g., URL, DOI if applicable)
- Restrictions for access and use, if any

Table 5-1 shows how provision of data product information relates to its grade for this sub-section of the quality assessment.

Table 3-1 – 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 missing.
Good	Some pieces of important information missing.
Excellent	Almost all required information available.
Ideal	All required information available.

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 standardised communications protocol
- The protocol is open, free and universally implementable
- The protocol allows for an authentication and authorisation 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**

- Metadata and data are richly described with a plurality of accurate and relevant attributes
- Metadata and data are released with a clear and accessible data usage license
- Metadata and data are associated with detailed provenance
- Metadata and data meet domain-relevant community standards

Table 5-2 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 5-2 – 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 licence.
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 licence.

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

Table 5-3 shows how a given EO data product should be graded for its format.

Table 5-3 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 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*.

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 Metadata and European Commission Joint Research Centre 2013), (INSPIRE Thematic Working Group Orthoimagery 2013), which they can be evaluated against.

Table 5-4describes how the assessment framework grades a products user documentation.

Table 5-4 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.

5.5 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.

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 2017c)). 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 5-5 shows how the assessment framework grades the metrological traceability documentation, based on its completeness.

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

Uncertainty Characterization

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.

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 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. Table 5-6 shows the uncertainty characterisation grading under the assessment framework.

Table 5-6 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.

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 5-7 shows how the ancillary data are graded under the assessment framework.

Table 5-7– 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.

5.6 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.

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, Varon, Irakulis-Loitxate, and Guanter 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 5-8 shows how the assessment framework grades the retrieval algorithm used to generate L2 products.

Table 5-8 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.

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 5-9 shows how geometric processing is graded.

Table 5-9 – 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.

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 5-10, and then a combined score determined.

Table 5-10 - Product Generation > Mission Specific Processing – Assessment Criteria

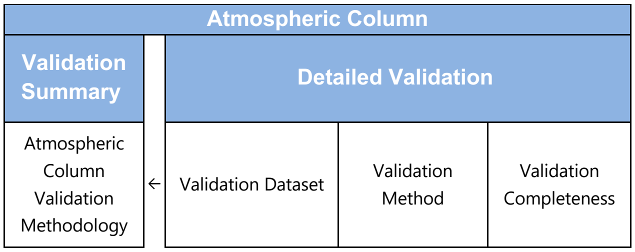
Grade	Criteria
Not Assessed	Assessment outside the scope of study.
Not Assessable	Additional processing steps not documented.
Basic	Additional processing steps documented. Additional processing steps not considered fit for stated purpose.
Good	Additional processing steps documented. All significant additional 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.

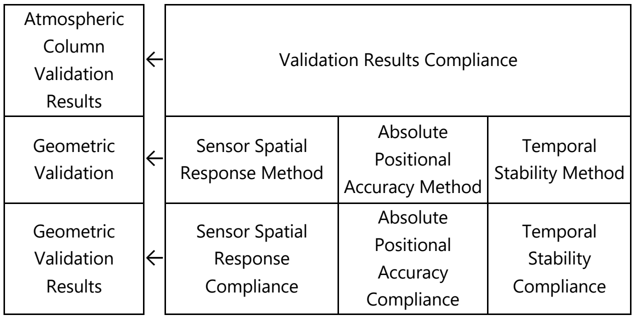
5.7 Detailed Validation

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 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 5-2.

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 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 5-5 – Detailed Validation Cal/Val Maturity Matrix and Validation Summary

5.8 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.

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 5-11 – 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.

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 traceable to SI

Table 5-12 shows how the validation approach is graded within the assessment framework.

Table 5-12 – 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)

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 5-13 shows how the validation completeness is graded within the assessment framework.

Table 5-13 – 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.

5.9 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 5-14 shows how the validation results are graded within the assessment framework.

Table 5-14 – 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.

5.10 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 centre. 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.

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.

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.

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.

5.11 Validation Summary

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

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

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.

5.12 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.

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.

6.0 Introduction: Quality assessment for reporting methane emissions

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 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.

Scope

This document 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 [RD-1] for this domain. Our quality assessment includes (1) traceability of the data to 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 6.2 provides a summary of the mission quality assessment framework. Section 6.3 provides a review of the atmospheric mission quality, as evidenced by its documentation. Section 6.4 provides guidelines for verifying the mission data quality is consistent with its stated performance. Section 6.5 describes the overall quality assessment guidelines which should result from the quality assessment from the previous sections.

6.1 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 [RD-2] and builds on the structure and reporting style developed in other similar work (e.g. [RD-3]). 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 5.3 and 5.4, 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 summarised 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.

6.2 Mission Data Supply Chain Assessment Overview

The product assessment outlined in this document forms part of a wider supply chain assessment summary (Figure 5-1). 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 1, 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 documentation 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 reproduceable from the L1 and L2 products and associated documentation (e.g. ATBD's and product descriptions).

6.3 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.

6.4 Product Evaluation Matrix

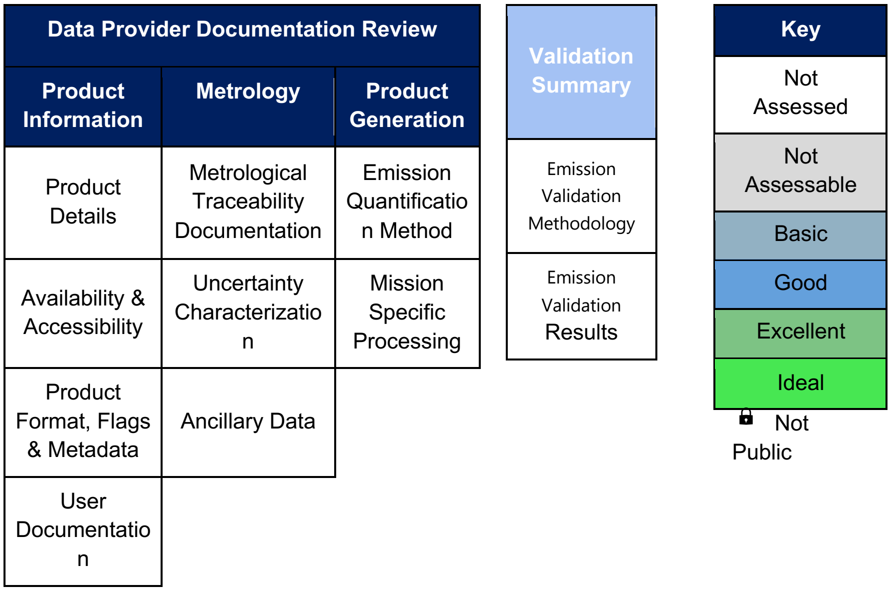
The product evaluation matrix provides a high-level color-coded summary of the quality assessment results. The matrix contains a column for each section of analysis, and cells for each subsection of analysis. Subsection grades are indicated by the color of the respective grid cell, which are defined 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.

6.4.1 Summary Product Evaluation Matrix

The *Summary Product Evaluation Matrix* is shown in Figure 6-1. 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	Emission Quantification Method
Availability & Accessibility	Uncertainty Characterization	Mission Specific Processing
Product Format, Flags & Metadata	Ancillary Data	
User Documentation		

Validation Summary
Emission Validation Methodology
Emission Validation Results

Key
Not Assessed
Not Assessable
Basic
Good
Excellent
Ideal

🔒 Not Public

Figure 6-1. Summary Product Evaluation Matrix.

Detailed Validation Maturity Matrix

The *Detailed Validation Maturity Matrix* (Figure 6-2) 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).

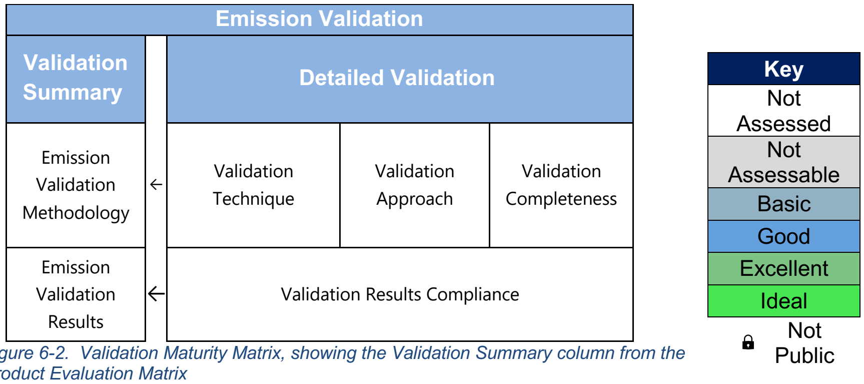


Figure 6-2. Validation Maturity Matrix, showing the Validation Summary column from the Product Evaluation Matrix

6.5 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.

6.6 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 6-1)*. This portion is shown in Figure 6-3.

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 6-3 – Data Provider Documentation Review Matrix

6.7 Product Information

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

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 organisation, including email address)
- Product access (e.g., URL, DOI if applicable)
- Restrictions for access and use, if any

Table 6-1shows how provision of data product information relates to its grade for this sub-section of the quality assessment.

Table 6-1– 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 missing.
Good	Some pieces of important information missing.
Excellent	Almost all required information available.
Ideal	All required information available.

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 [RD-4], 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 standardised communications protocol
- The protocol is open, free and universally implementable
- The protocol allows for an authentication and authorisation 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**

- Metadata and data are richly described with a plurality of accurate and relevant attributes
- Metadata and data are released with a clear and accessible data usage license
- Metadata and data are associated with detailed provenance
- Metadata and data meet domain-relevant community standards

Table 6-2 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 6-2– 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 licence.
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 licence.

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 [RD-5] 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 [RD-6], or, for data from the European Union, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive [RD-7]
- 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 6-3 shows how a given EO data product should be graded for its format.

Table 6-3 – 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 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*.

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 [RD-8], [RD-9], which they can be evaluated against.

Table 6-4 describes how the assessment framework grades a products user documentation.

Table 6-4– 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.

6.8 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.

Metrological Traceability Documentation

Traceability is defined in the vocabulary of metrology (VIM) [RD-10] 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 [RD-11]). 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 centred “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 [RD-12].

Table 6-5 shows how the assessment framework grades the metrological traceability documentation, based on its completeness.

Table 6-5 – Metrology > Metrological Traceability Documentation – Assessment Criteria

6.8.1 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) [RD-13].

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 [RD-14]. There have also been some initiatives, like the previously mentioned FIDUCEO project, that have applied metrology to historical sensor data records [RD-15].

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

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.

Table 6-6 – 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.

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 6-7 shows how the ancillary data are graded under the assessment framework.

Table 6-7 – 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.

6.9 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.

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 [RD-35] 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 6-8 shows how the assessment framework grades the retrieval algorithm used to generate L2 products.

Table 6-8 – 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.

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 Tab le 6-10, and then a combined score determined.

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.

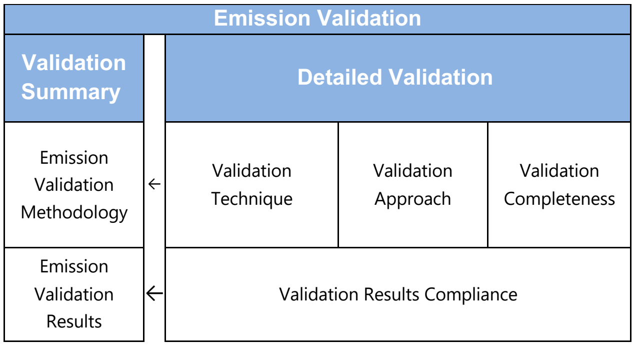
Table 6-10 - Product Generation > Mission Specific Processing – Assessment Criteria

6.10 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 5), 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 6-1.

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 6-4 – Detailed Validation Cal/Val Maturity Matrix and Validation Summary

6.10.1 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.

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 6-11 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 6-11– 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.

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 [RD-33]) are required for the highest assessment grades.

Table 6-12 shows how the validation approach is graded within the assessment framework.

Table 6-12– 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)
Ideal	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)

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 6-13 shows how the validation completeness is graded within the assessment framework.

Table 6-13– 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 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.

6.10.2 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 6-14shows how the validation results are graded within the assessment framework.

Table 6-14– 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.

6.10.3 Validation Summary

The *Validation Summary* provides a synthesis of the per performance metric assessments provided in the *Detailed Validation Cal/Val Maturity Matrix* (Figure 6-3). 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.

6.11 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.

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 defining 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.

Total Carbon Column Observing Network (TCCON)

Description

TCCON has been a longstanding tool for validating satellite GHG column data, such as CO_2 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 ~-.02 cm^{-1}. These instruments retrieve total column amounts of CO_2, CH_4, N_2O, 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

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

Variable depending on individual validation activity (site, satellite, and time dependent). Good TCCON column retrieval uncertainties e.g., <1 ppm for CO2, <5 ppb for CH4 (1σ) (Wunch et al. 2011).

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

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

Minimal bias and long-term drift relative to TCCON. 2σ uncertainties of 0.6 ppm for CO_2 and 2.2 ppb 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 with in 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 CO2 profiles and satellite retrieved CO2 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 (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 formalised 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 utilised as supplementary validation datasets in support of more mature, widespread ground-based validation networks.

Scope of Representativeness

Some degree of mismatch between aircraft/sonde profile and satellite column, as validation data must be extrapolated to match full vertical atmospheric column. Limited spatial and temporal coverage as studies often performed on a case-by-case basis.

Quality

Variable depending on in situ technique, specific study, etc. Example 1σ in situ instrument precisions: 0.02 ppm for CO2, 0.5 ppb for CH4 (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
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
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
GUM	Guide to the Expression of Uncertainty in Measurements
L1	Level 1
L2	Level 2
LSF	Line Spread Function
MTF	Modulation Transfer Function
NASA	National Aeronautics and Space Administration
NetCDF	Network Common Data Format
NPL	National Physical Laboratory, UK
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)
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

References:

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.

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

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

Chander, G., Hewison, T. J., Fox, N. P., Wu, X., Xiong, X., and Blackwell, W. J. (2013), Overview of Intercalibration of Satellite Instruments, *IEEE Trans. Geosci. Remote Sens.*, 51(3), 1056-1080, doi:10.1109/TGRS.2012.2228654.

COP28 UAE (2023), Oil and Gas Decarbonization Charter Launched to Accelerate Climate Action, Available at: https://www.cop28.com/en/news/2023/12/Oil-Gas-Decarbonization-Charter-launched-to--accelerate-climate-action.

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.

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.

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.

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

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

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.

Fougnie, B., and Bach, R. (2009), Monitoring of Radiometric Sensitivity Changes of Space Sensors Using Deep Convective Clouds: Operational Application to PARASOL, *IEEE Trans. Geosci. Remote Sens.*, 47(3), 851-861, doi:10.1109/TGRS.2008.2005634.

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.

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.

Global Methane Pledge (2021), Available at: https://www.globalmethanepledge.org.

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

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.

Govaerts, Y. M., Rüthrich, F., John, V., Quast, R., and John, V. O. (2018), Climate Data Records from Meteosat First Generation Part I: Simulation of Accurate Top-of-Atmosphere Spectral Radiance over Pseudo-Invariant Calibration Sites for the Retrieval of the In-Flight Visible Spectral Response, *Remote Sens.*, 10(12), 1959, doi:10.3390/rs10121959.

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.

Holl, G., Woolliams, E., and Mittaz, J. (2019), FIDUCEO D2.2 (HIRS): Report on the HIRS FCDR Uncertainty, Available at: https://cordis.europa.eu/project/id/638822/results.

Hulswar, S., Soni, V. K., Sapate, J. P., More, R. S., and Mahajan, A. S. (2020), Validation of satellite retrieved ozone profiles using in-situ ozonesonde observations over the Indian Antarctic station, Bharati, *Polar Sci.*, 25, 100547, doi:10.1016/j.polar.2020.100547.

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

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.

International Energy Agency (IEA) (2024), EU Methane Regulations, Available at: https://www.iea.org/policies/18209-eu-methane-regulations#.

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.

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.

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

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.

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.

Karion, A., Sweeney, C., Tans, P., and Newberger, T. (2010), AirCore: An Innovative Atmospheric Sampling System, *J. Atmos. Oceanic Technol.*, doi:10.1175/2010JTECHA1448.1.

Kunkel, K. et al. (2023), https://pubs.acs.org/doi/10.1021/acs.est.3c00229.

Lyapustin, A., et al. (2014), Scientific impact of MODIS C5 calibration degradation and C6+ improvements, *Atmos. Meas. Tech.*, 7(12), 4353-4365, doi:10.5194/amt-7-4353-2014.

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.

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.

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/rs13061234.

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.

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.

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

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

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.

Rüthrich, F., Woolliams, E., Govaerts, Y., Quast, R., and Mittaz, J. (2019), FIDUCEO D2.2 (MVIRI): Report on the MVIRI FCDR Uncertainty, Available at: https://cordis.europa.eu/project/id/638822/results.

Scanlon, T. (2017), QA4ECV Product Documentation Guidance: Algorithm Theoretical Basis Document, Available at: Site archive - KNMI (sitearchief.nl).

Scanlon, T. (2017), QA4ECV Product Documentation Guidance: Product User Manual, Available at: Site archive - KNMI (sitearchief.nl).

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

Scanlon, T. (2017), QA4ECV Product Documentation Guidance: Validation and Intercomparison Report, Available at: http://www.qa4ecv.eu/sites/default/files/QA4ECV%20Validation%20Guidance.pdf.

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.

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

Stone, T. C., Kieffer, H., Lukashin, C., and Turpie, K. (2020), The moon as a climate-quality radiometric calibration reference, *Remote Sens.*, 12(11), 1–17, doi:10.3390/rs12111837.

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.

Taylor, M., Mittaz, J., Desmons, M., and Woolliams, E. (2019), FIDUCEO D2.2 (AVHRR): Report on the AVHRR FCDR Uncertainty, Available at: https://cordis.europa.eu/project/id/638822/results.

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.

Thome, K., Smith, N., and Scott, K. (2001), Vicarious calibration of MODIS using Railroad Valley Playa, in *IGARSS 2001: Scanning the Present and Resolving the Future*, IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217), 3, 1209–1211, doi:10.1109/IGARSS.2001.976794.

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

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.

Vermote, E., Santer, R., Deschamps, P. Y., and Herman, M. (1992), In-flight calibration of large field of view sensors at short wavelengths using Rayleigh scattering, *Int. J. Remote Sens.*, 13(18), 3409–3429, doi:10.1080/01431169208904131.

White House (2022), Executive Order on the Implementation of the Energy and Infrastructure Provisions of the Inflation Reduction Act, Available at: https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-the-implementation-of-the-energy-and-infrastructure-provisions-of-the-inflation-reduction-act-of-2022/.

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.

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.

World Meteorological Organization (WMO) (2024), Global Greenhouse Gas Watch (G3W), Available at: https://wmo.int/activities/global-greenhouse-gas-watch-g3w.

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

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

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