White Paper

Validation and Metrics for Emissions Detection by Satellite

Jason McKeever and Dylan Jervis

September 2022

1 CONTENTS

2 INTRODUCTION

Detecting and quantifying greenhouse gas emissions from individual sites by satellite remote sensing has emerged as a powerful new method in recent years. As more and more players enter the field, based on a variety of technologies for both instrumentation and data processing, there is a need for standardized methods for evaluating the performance of these systems. This document is focused on the specific example of satellite-based methane detection – but the principles can easily be applied to detection of other gases, and to detection from different platforms (such as aircraft).

These measuring systems must be validated using the statistical metrics of *sensitivity*, *specificity* and *quantification accuracy*. The sensitivity tells us how effective the system is at detecting an emission that is truly present (the rate of *true positives*). The sensitivity increases for higher (true) emissions rates, and approaches zero for very low rates – this leads to the important concept of a *detection limit*. The specificity tells us how good we are at suppressing *false positives* – declared detections when no emission of the gas of interest is truly occurring. False positives are highly undesirable by the end-users, including industrial operators seeking to monitor and manage their emissions – so there is typically a strong requirement for high specificity. Finally, the quantification accuracy must be measured and specified to give credibility to emission rates and their uncertainties reported by the measurement provider.

An effective emissions detection system has the following requirements, all of which must be satisfied simultaneously.

(a) a well-characterized detection limit based on a Probability of Detection curve
(b) a well-characterized, high specificity (low rate of false positives)
(c) well-characterized quantification accuracy

In the sections that follow, we elaborate on all these concepts with concrete examples of how they should be implemented in practice.

3 SENSITIVITY AND DETECTION LIMIT

The statistical concept of sensitivity expresses the likelihood that a truly occurring emission is positively detected. In some cases (such as detecting if a person is infected with a certain virus), we are simply interested in the presence or absence of a substance – and the sensitivity of a test can be captured in a single percentage. For emissions detection, however, lower emission rates are more difficult to detect – so the sensitivity can be expressed as an increasing function of emission rate (Figure 1). The concept of *detection limit* flows naturally from this – what is the lowest detectable emission rate for a given system? This "detection limit" is also known by various other terms including "limit of detection", "detection threshold" and "minimum detectable emission".

3.1 Definition of the Detection Limit

The starting point for determining the detection limit for an emissions measurement system is a Probability of Detection (PoD) curve, such as the theoretical example in Figure 1. As mentioned above, this curve is not a "step-function" – rather, there is a smooth increase in detection probability as the emission rate increases. The "soft" nature of the threshold arises due to randomness of measurement noise, and uncontrolled environmental variables such as turbulence, wind conditions and terrain reflectance. Due to its strong influence on the detection probability, the PoD curve should be expressed as a function of wind speed u (the example in Figure 1 could pertain to a single value of u). In practice, as we describe below, this type of curve must ideally be inferred by fitting a model to empirical detection results (simulations can be used for the case of pre-launch systems).

As for how to quote a single value for detection limit, GHGSat typically quotes the rate Q_{50} at which PoD = 50%, for a reference wind speed of u_0 = 3 m/s. Other threshold probability values such as 90% or 95% are sometimes used - Q_{50} and Q_{90} are shown in Figure 1. We believe this proposed definition is in accordance with best practices documented in the scientific literature.

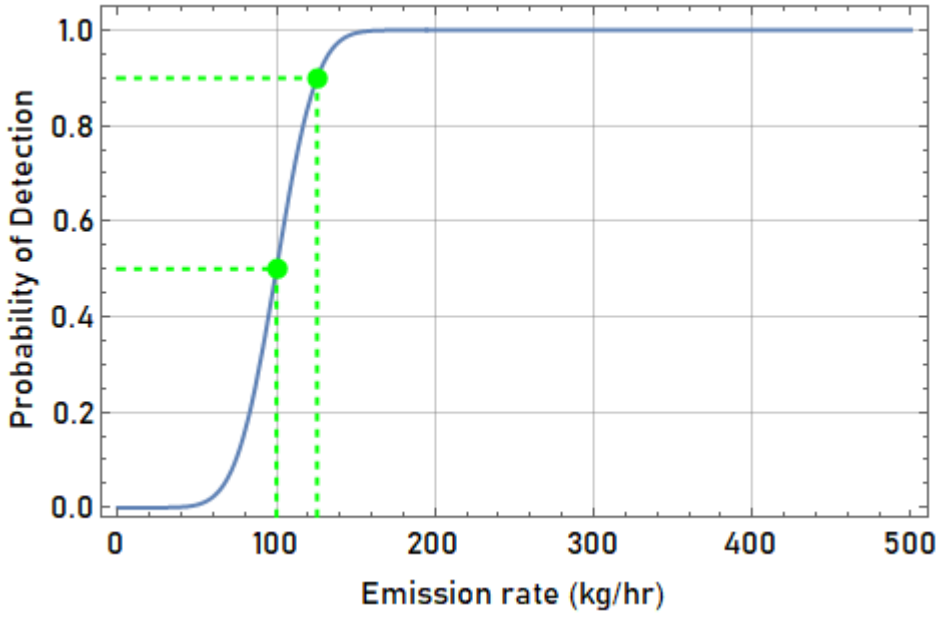

Figure 1: A simulated example of a Probability of Detection curve (not inferred from a real measurement system). Dashed lines indicate the (50%, 90%) detection thresholds (Q_{50}, Q_{90}) = (100 kg/hr, 126 kg/hr).

In the following sections we describe several ways detection limit can be evaluated. Each of these approaches has its own strengths and limitations, and therefore each can be appropriate depending on the specifics of the measurement being done.

3.2 IMPORTANT FACTS ABOUT THE DETECTION LIMIT

There are several points about the detection limit that are common to all known remote methane sensing platforms.

1. The detection limit is probabilistic

   There is no "hard cutoff" emission rate defining the boundary between successful and unsuccessful detection. Rather, there is a curve expressing probability of detection (PoD) vs emission rate (Figure 1).

2. The detection limit depends on wind speed

   The detailed shape of an emission plume is determined by complex meteorological factors. However, higher wind speed leads to increased dilution (lower methane densities) in the plume. This constitutes a signal reduction for all detection systems and therefore the detection limit increases with wind speed. A simple, frequently used approximation is described in Jacob et al., 2016, though one empirical analysis suggests a somewhat softer dependence (Johnson et al., 2021 and Figure 3).

3. The detection limit is condition-dependent

   All methane sensing systems based on infrared spectroscopy (including LIDAR systems) rely on detecting light backscattered from surfaces beneath the emissions plume. The properties of these surfaces affect the detection limit. Higher surface reflectance (e.g., from drier terrain classes) boosts the optical signal and leads to lower detection limits. Furthermore, many systems are subject to spurious signals when the surface is highly non-uniform - in other words, detection is easier over near-uniform, desert-like scenes.

4. Point source vs area source

   A point source is typically defined as being smaller than the spatial resolution (ground sampling distance, or GSD) of the detection system. These have the lowest detection limit because they give the highest densities near the source. All else being equal, diffuse area-emitters larger than the GSD have higher detection limits owing to the associated density reduction.

5. Prior knowledge matters (targeted facilities vs area surveys)

   Unambiguous plume detection depends on the ability of a software algorithm or a human technician to confidently discriminate between a plume signal and noise/errors in the underlying data. If the objective is to image a *targeted facility* and look for an emissions plume, this is an advantage, since only one point of origin must be considered. This means that surveying an area where *leaks could be coming from anywhere (or from any one of multiple facilities)* is more challenging and has higher detection limits than measuring targeted facilities.

3.3 METHODS FOR EVALUATING THE DETECTION LIMIT

3.3.1 Detection limit as smallest detected emission

The detection limit is sometimes specified as the smallest emission that has ever been detected by a given instrument. **By this measure, GHGSat's satellite detection limit would be 42 kg/hr** for the example seen in Figure 2, or even lower if estimated emission rates with very large uncertainty were to be considered. This is a very simple approach, but it has several shortcomings, beyond the fact that it does not conform to the definition proposed above (rate for which PoD curve equals a standardized value, such as Q_{50}).

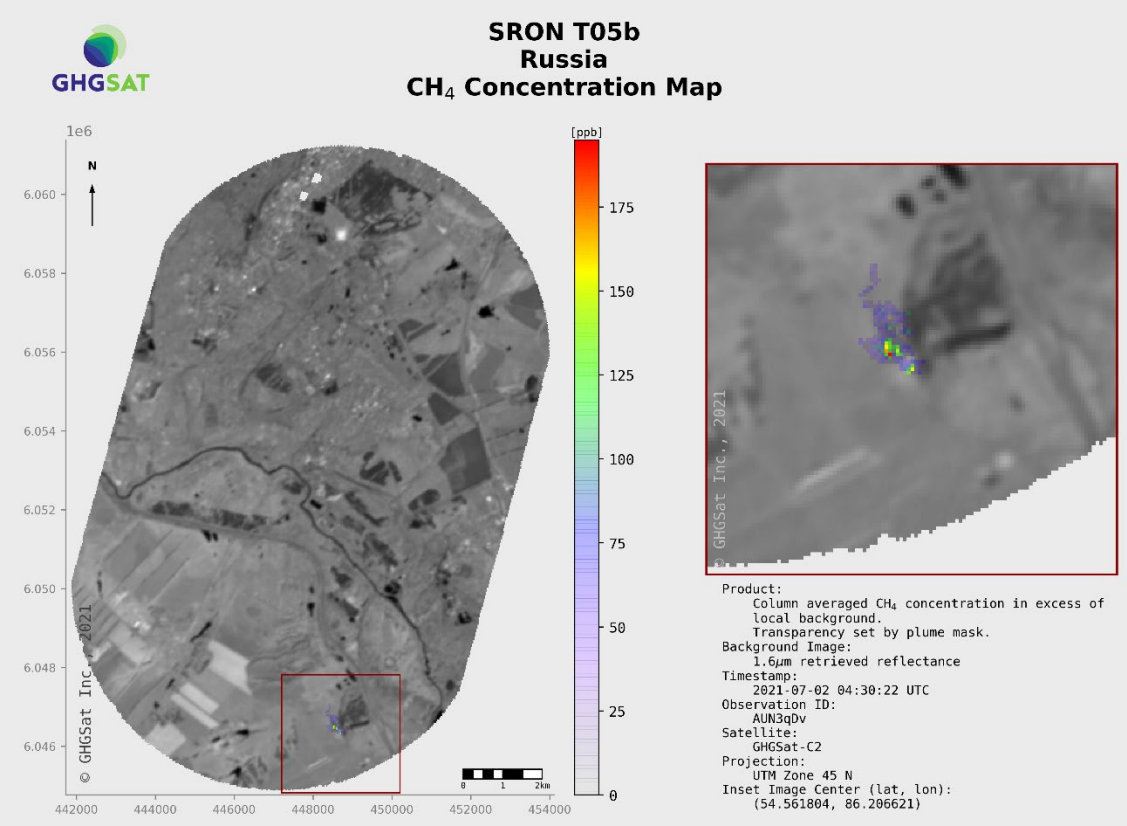

Figure 2: The plume shown in this observation had a retrieved source rate of 42 kg/hr.

The primary issue is that it is susceptible to the "cherry-picking" of results. For instance, detecting a plume in a time and place with lower wind speeds and bright, homogenous terrain is generally easier than over dark, heterogeneous terrain in windy conditions. Furthermore, this amounts to sampling from the low-rate "tail" of the distribution. The more measurements are done, the more events there are in the tail, so that the lowest detected rates will be lower if the number of measurements is higher.

Another issue is that the true release rate is unknown *a priori*, so the result is subject to quantification error in inferring the source rate from the measured plume (see Section 5).

3.3.2 Detection limit from controlled releases

Another method is to define the detection limit from "controlled-release" measurement campaigns. These consist of a series of measurements in which methane is emitted at a known, adjustable rate.

- When a series of controlled releases is done to assess the detection limit, they are typically done from a fixed release point – this means the measurement team has knowledge of the release point. Consequently, this method is most appropriate for the use case of measuring *targeted facilities*.
- If the releases are organized and conducted independently of the measurement team ("single-blind" campaign), this helps third parties gain confidence in the validity of the results, both in terms of detection and quantification. In a single-blind study, the measuring team is not aware of the true release rates until after their rate analysis is completed and delivered to the organizers.
- Another approach is for the measurement provider to be blind to the release location as well (Johnson et al., 2021). This is challenging to achieve in practice given that the release infrastructure must be moved around, ideally to locations representative of real-world leaks. If the measurement provider really is fully blind to the location, magnitude (and even existence) of the release, this becomes a more powerful technique for characterizing detection limit.
- Ideally, many releases are done over a range of rates above and below the detection limit, so that a full detection probability curve can be built up with adequate statistics.

For the case of detections by satellite, it can be a challenge to get enough measurements to do a proper statistical inference of the detection limit, since relatively large rates are required, and repeated measurements are limited by the frequency of satellite overpasses. However, at GHGSat we are committed to pursuing this key validation method. We have participated in several controlled-release campaigns, both internally organized and single-blind with third parties. This includes an independently run single-blind controlled release campaign by Stanford University where we successfully detected and quantified all methane emission rates (Sherwin et al., 2022).

The detection results from these releases (through August 2022) are captured in Figure 3. The results are fit to a model with the truth rate and wind speed as independent variables. The model allows for a "soft" transition from zero to 100% detection probability, with the 50% threshold depending linearly on wind speed. In other words, it gives us a wind-speed-dependent PoD curve. At a typical wind speed of 3 m/s, the model implies **Q_{50} = 117 kg/hr**.

As mentioned above, the detection probability also depends on environmental variables besides the wind speed. For example, it should be easier to detect plumes when the instrument collects more light to make its measurement (i.e. over bright scenes at noontime in summer - when the surface albedo is greater and the sun is higher in the sky). In practice, it is difficult to conduct enough controlled-release observations under varied enough conditions to empirically determine the dependence of the detection probability on each environmental variable. In lieu of empirically inferred dependencies, we can provide a physics-based estimate (Appendix A). For the GHGSat releases in Figure 3, the mean albedo and solar angles were very close to the corresponding global mean across all our "real-world" observations. Using these mean values in the Appendix A model, we estimate that our inferred detection limit is only

slightly affected (biased by less than 3%). However, if releases were conducted in highly favourable conditions (for example albedo of 0.5) – the inferred detection limit could be biased low by over 25%.

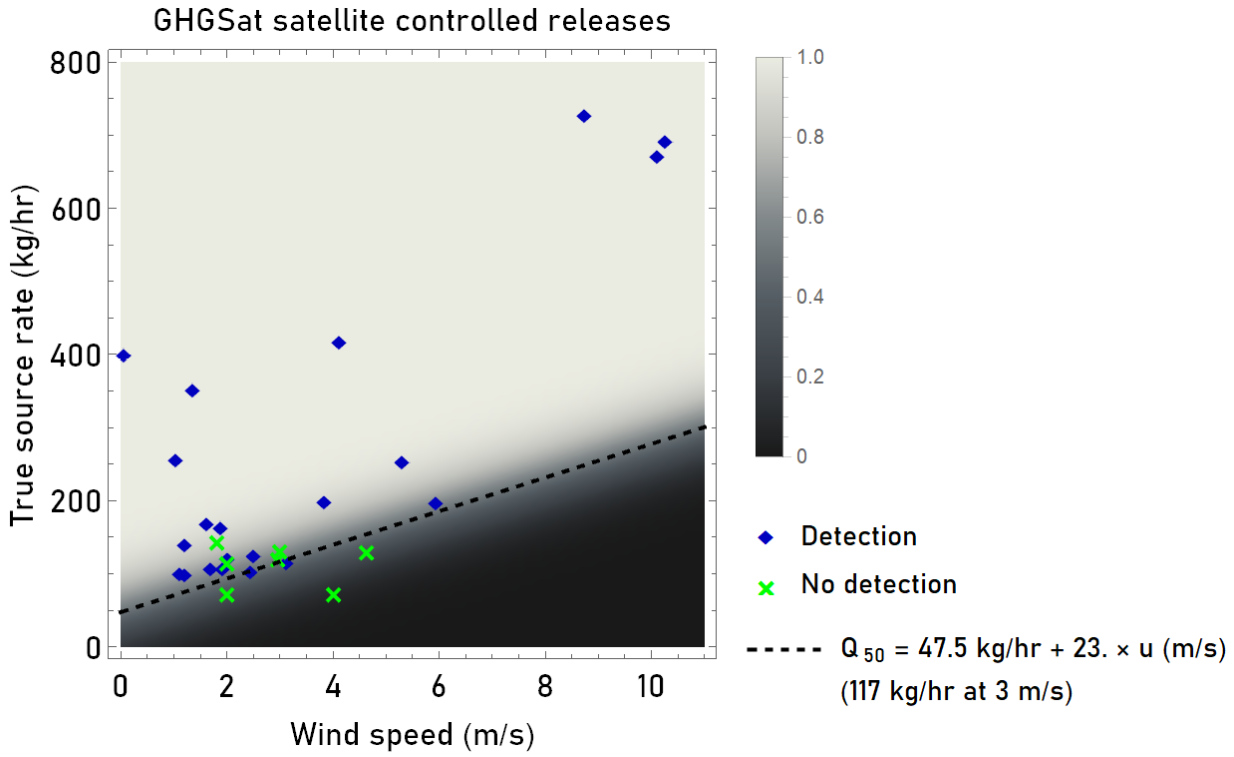

Figure 3: Regression analysis of detection limit from controlled releases measured by GHGSat's satellites (Sep 2020 - Aug 2022). The gray scale represents the probability of detection (colour bar). The dashed line gives the best-fit threshold Q_{50} at each value of u.

3.3.3 Detection limit from aggregate field data

A method that infers detection limit from an analysis of all detected emission rates avoids many of the problems of the previously mentioned methods: the data used includes thousands of plumes, all measured "in the wild" all over the world, over all kinds of environmental conditions (time of year, wind conditions, etc.), and in representative operational conditions. By this last point we mean that whereas a human technician or plume detection algorithm may be especially successful in detecting plumes from "high-priority" controlled-release observations, their performance in more nominal situations could be somewhat different.

This analysis method assumes that the true probability of finding a plume at a given emission rate follows a well-established power-law mathematical function (Ehret et al., 2021; Lauvaux et al., 2021). Any deviation from this function at lower emission rates is attributed to the instrument's detection limit.

While this method is perhaps the most holistic measure of an instrument's operational detection limit, it has two important shortcomings:

- *It is vulnerable to source rate quantification error*. There is no "ground truth" source rate to compare against in this method. Instead, the estimated source rates are assumed to be "truth". Therefore, if there is a systematic bias in the source rate estimate, the inferred detection limit will be similarly biased. See Section 5 for further discussion of quantification accuracy and bias.
- *It is vulnerable to false positives*. Another assumption behind this method is that it only includes detections that have been rigorously vetted as being "true" positives. Any false positives, which are vastly more likely to occur at low emission rates, will bias this estimate low since additional low-emission detections would be erroneously included in the aggregate field dataset. See section 4 for further discussion of false positives and specificity.

Therefore, one should only place confidence in detection limits derived from an aggregate field analysis when they are accompanied by a full validation of the source rate quantification (Section 5), a detailed error estimate and a specificity analysis (Section 4).

With the launch of the latest satellites in the GHGSat constellation, GHGSat is collecting up-to-date aggregate emissions data that we intend to analyze – and infer the detection limit from - in a future peer-reviewed publication.

3.3.4 Detection limit from column density precision and spatial resolution

A less direct way to define the detection limit is to infer it from the measured "column density" precision and the spatial resolution of instrument. This method is often used to estimate the performance of prospective future instruments for which no measurement data yet exists.

This method often involves placing simulated plumes in simulated background noise fields that have been generated using the assumed spatial resolution and measurement precision of the prospective instrument. An emission rate detection threshold is then inferred from heuristic arguments such as, for example, a detected plume must have a peak column density enhancement that is twice the standard deviation of the noise field (Jacob et al., 2016).

Using this method with GHGSat's *measured* **1.6% of background** methane column density precision and **27 m** spatial resolution as input we would infer a detection limit of **~50 kg/hr.**

Because this method does not use any measured emission rate data, it is susceptible to errors in the model used to translate column density precision and spatial resolution into emission rate detection probability. Moreover, for the case of prospective instruments, optimistic projections for the column density precision and spatial resolution would bias this detection limit estimate low. Realistic estimates for the character of an instrument's measurement error are difficult to predict before real data is available.

3.3.5 Detection limit: Summary of methods and conclusion

The most robust estimates of a technology's detection limit derive from an analysis of either **controlled-release** or **aggregate field** measurements:

- The detection limit estimated from **controlled-release** data is applicable when measuring emissions from targeted facilities.
- The detection limit estimated from **aggregate field** data is applicable to wide area emissions surveys where the emitting source locations are unknown.

The **lowest rate detected** can be an eye-catching performance indicator, but it is clearly not a statistically representative estimate of performance. Inferring the detection limit from the **column density precision** seems most useful for prospective instruments where real measurement data is not yet available. It should be noted that these latter two methods tend to underestimate the detection limit compared to the statistical analysis of the controlled-release and aggregate field data methods.

Table 1 compares the features of the different methods described and lists the estimated GHGSat detection limit for each one.

Method	Estimated GHGSat detection limit	Statistical method, gives PoD curve	No assumption for true rate distribution	No bias from rate quantification error	Covers full range of field conditions	Based on actual detected plumes	Accounts for false positives
Lowest rate detected	42 kg/hr	☒	☑	☒	☒	☑	☒
Controlled releases	117 kg/hr*	☑	☑	☑	☒	☑	☑
Aggregate field statistics	TBD	☑	☒	☒	☑	☑	☒
Column noise inference	50 kg/hr	☒	☑	☑	☒	☒	☒

*Table 1: Advantages of the described methods for inferring detection limit. *Q50 at u=3 m/s.*

3.4 Automated Plume Detection

One important factor influencing the detection limit is the method used to determine the presence of a plume. As the volume of satellite data increases, it becomes impractical for human analysts to do this by inspection. Therefore, there is a growing interest in detecting plumes by artificial intelligence (AI) or other automated techniques. However, because these techniques typically do not account for the various causes and presentations of systematic measurement errors as well as human operations technicians do, the results must be carefully validated. This is especially the case when these automated techniques are required to balance sensitivity (increasing the rate of true-positives) with specificity (reducing the rate of false-positives).

AI detection techniques require training an algorithm – often a convolutional neural network – on hundreds or thousands of simulated or human-detected plumes (Jongaramrungruang et al., 2022). This AI algorithm is then sometimes supplemented with additional information about aerosol concentration, albedo, or other components of the remote sensing measurement that are correlated with methane measurement errors. Non-AI automated techniques often require that, for example, the plume's column density enhancement must be at least a certain multiple of the background noise level and spatially correlated to some degree (Duren et al., 2019).

It should be noted that AI and other automated plume detection techniques are subject to the same physical constraints of the remote sensing measurement as human or other plume detections methods. Therefore, while it is possible that a well-designed AI system could outperform human analysts and yield an improved detection limit, it is unlikely that the AI system can detect *dramatically* lower rates.

When a detection system is fully validated through a detection limit and specificity analysis, we are validating not only the instrument performance, but any such detection algorithm as well. As mentioned above, in the case of these AI methods, we must be especially vigilant about specificity, since a poorly "tuned" (too aggressive) algorithm could appear to have artificially low detection limit due to a high rate of false positives. Ideally, such a system should be fully evaluated by a third party using single-blind controlled releases where the detection team is not aware of timing, location or rate of the releases, with results published in the peer-reviewed literature. Alternatively, the system could be validated against an independent well-established measuring system (not necessarily satellite-based). As of this writing, we are not aware of any such published work to fully evaluate a detection system including automated plume detection software.

4 SPECIFICITY AND FALSE POSITIVES

A "false positive" is a declared detection event for which there is no true underlying emission. This is typically caused by spurious signals in the data. These events are highly undesirable from the perspective of the industrial operators and other end-users of the satellite data.

Therefore, we require a metric known as *specificity,* quantified using the ratio

$$\frac{TN}{TN + FP}$$

- where TN is the number of *true negatives* – events where no detection is declared and there is no true emission above the satellite's detection limit
- FP is the number of *false positives* – events where a detection is declared (erroneously) and there is no true emission above the satellite's detection limit.
- Importantly, these statistics only include events for which it is known that no true emission was occurring at the time of measurement.
- In the ideal case, the specificity approaches unity.

Quantifying the specificity rigorously for a real satellite measuring system is challenging for the following reasons:

- In order to accumulate good statistics over a range of conditions, a reasonable number of sites must be measured repeatedly for which the true emission "state" is known at the time of satellite measurement.
- Among these measurements, a reasonable number must be done with no true emissions
- The sites must either be controlled release facilities, or sites for which methane emissions are continuously monitored by an independent measuring technology.
- Ideally, the team analyzing the (non-)detection events should be blind to the true emission rate.
- Unfortunately, data from routine "field operations" where the true emission state is unknown cannot be used to quantify this metric.

As of this writing, GHGSat does not have sufficient sample size to quote a statistically meaningful value for specificity for our satellites. However,

(a) We have never had a false positive from a single-blind controlled release with no true emission (although our sample size of these events is very limited to date)
(b) The feedback we receive from end users of our data very rarely includes cases of suspected false positives.

In order to mitigate our false-positive rate, GHGSat 1) includes a robust error estimate for every ground cell in our methane retrieval field to flag spurious false-positive signals whenever the error exceeds a certain threshold, 2) understands how disruptive a false-positive report to a client can be and so maintains rigorous quality assurance in our human plume detection procedure.

Finally, as a general statement, there is often a trade-off between improving the detection limit and reducing the false-positive rate (increasing specificity). As we have touched on in preceding sections, this can lead to a situation where the detection limit is pushed artificially low due to an increase in false positives. Therefore, the best practice for measurement providers is to specify the specificity (or false positive rate) in addition to the detection limit.

5 QUANTIFICATION ACCURACY

Once a plume has been detected, the next step is to quantify or "retrieve" the underlying emission rate, also known as source rate. There are many factors that influence the accuracy of such an estimate, and measurement providers must account for all such known factors and produce an uncertainty or error estimate along with the retrieved rate.

While the details of these error sources and the methods for source rate estimation are beyond the scope of this white paper, it is important to describe how the accuracy should be assessed. Best practice from the scientific literature is to perform a series of single-blind controlled releases over a large range of rates above the detection limit. The measurement provider estimates the emission without knowledge of the true rate, and the release organizers perform the analysis to compare true rates to retrieved rates. This has been done in several studies for methane sensing from aircraft (Johnson et al., 2021; Ravikumar et al., 2019; Sherwin et al., 2020), and there is one recent study for satellites – in which GHGSat was a participant (Sherwin et al., 2022).

GHGSat's results from this study are shown in Figure 5, with an example of a regression analysis showing a relatively small bias – the slope deviates from unity by 2.7%, and there is no evidence of nonlinearity (we have plans to increase the sample size in coming months). This does not imply that an individual plume can be quantified with 2.7% accuracy (as evidenced by the error bars on each point). Rather, this type of analysis tells us about the size and magnitude of any systematic bias, miscalibration or nonlinearity in the quantification. While each observation of a plume has relatively large random error (over 20% in Figure 5), this random uncertainty can in principle be reduced if we could measure the same plume repeatedly and average the results. Systematic bias, however, does not "average down" – rather it must be characterized by this type of analysis.

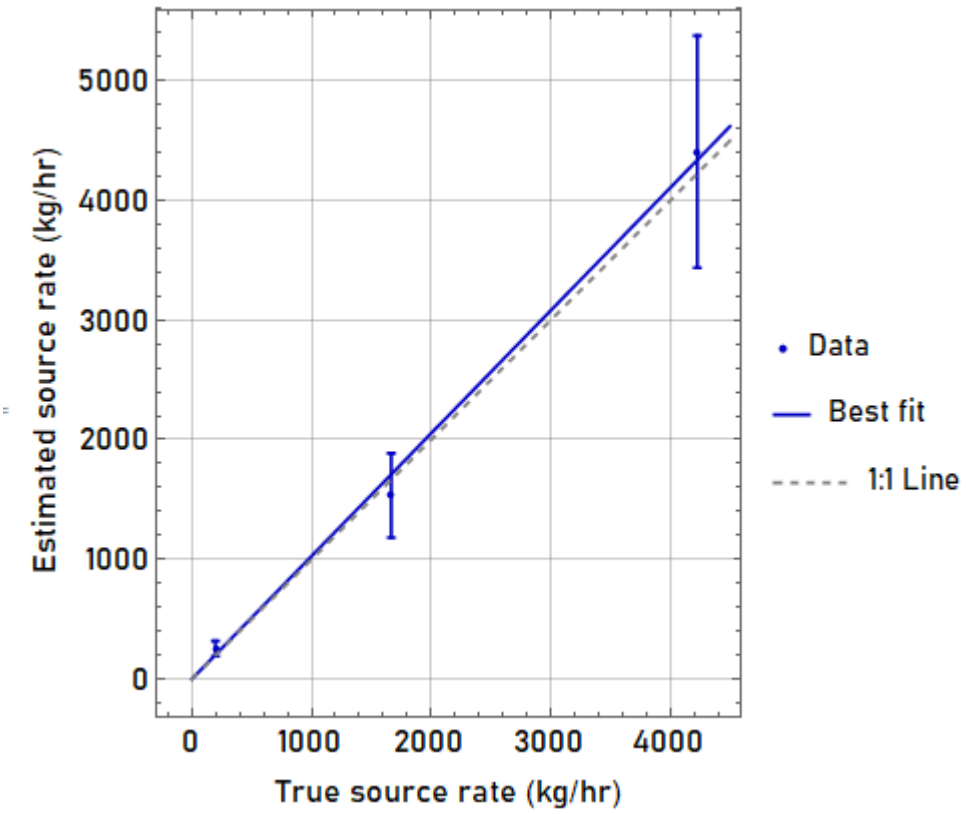

Figure 4: GHGSat parity chart from stage 2 unblinding in (Sherwin et al., 2022). True source rates are mean flow meter readings from 300 seconds prior to the measurement. Fit shown here is a simple linear regression forced through the origin – the resulting slope is 1.027.

6 SUMMARY/CONCLUSIONS

In summary, rigorous validation of an emissions measurement system is comprised of three critical metrics:

- Detection limit from a PoD curve
- Specificity
- Quantification accuracy

These metrics are interdependent – for example, poor specificity will compromise the validity of a detection limit analysis by artificially increasing the number of low-rate detection events.

7 REFERENCES

Duren, R. M., Thorpe, A. K., Foster, K. T., Rafiq, T., Hopkins, F. M., Yadav, V., Bue, B. D., Thompson, D. R., Conley, S., Colombi, N. K., & others. (2019). California's methane super-emitters. *Nature*, *575*(7781), 180–184.

Ehret, T., de Truchis, A., Mazzolini, M., Morel, J.-M., Lauvaux, T., Duren, R., Cusworth, D., & Facciolo, G. (2021). *Global Tracking and Quantification of Oil and Gas Methane Emissions from Recurrent Sentinel-2 Imagery*.

Jacob, D. J., Turner, A. J., Maasakkers, J. D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., & Frankenberg, C. (2016). Satellite observations of atmospheric methane and their value for quantifying methane emissions. *Atmospheric Chemistry and Physics*, *16*(22), 14371–14396.

Johnson, M. R., Tyner, D. R., & Szekeres, A. J. (2021). Blinded evaluation of airborne methane source detection using Bridger Photonics LiDAR. *Remote Sensing of Environment*, *259*, 112418. https://doi.org/10.1016/J.RSE.2021.112418

Jongaramrungruang, S., Thorpe, A. K., Matheou, G., & Frankenberg, C. (2022). MethaNet – An AI-driven approach to quantifying methane point-source emission from high-resolution 2-D plume imagery. *Remote Sensing of Environment*, *269*, 112809. https://doi.org/10.1016/J.RSE.2021.112809

Lauvaux, T., Giron, C., Mazzolini, M., d'Aspremont, A., Duren, R., Cusworth, D., Shindell, D., & Ciais, P. (2021). Global Assessment of Oil and Gas Methane Ultra-Emitters. *ArXiv Preprint ArXiv:2105.06387*.

Ravikumar, A. P., Sreedhara, S., Wang, J., Englander, J., Roda-Stuart, D., Bell, C., Zimmerle, D., Lyon, D., Mogstad, I., Ratner, B., & Brandt, A. R. (2019). Single-blind inter-comparison of methane detection technologies – results from the Stanford/EDF Mobile Monitoring Challenge. *Elementa*, *7*. https://doi.org/10.1525/elementa.373

Sherwin, E. D., Chen, Y., Ravikumar, A., & Brandt, A. R. (2020). *Single-blind test of airplane-based hyperspectral methane detection via controlled releases*.

Sherwin, E. D., Rutherford, J. S., Chen, Y., Aminfard, S., Kort, E. A., Jackson, R. B., & Brandt, A. R. (2022). *Title: Single-blind validation of space-based point-source methane emissions detection and quantification*.

APPENDIX A: VARIATION OF DETECTION LIMIT WITH ALBEDO AND SOLAR ZENITH ANGLE

We expect the 50% detection limit Q_{50} to scale inversely with measurement precision σ: $Q_{50} \propto \sigma$, which is limited by random noise at the level of the camera pixels. More specifically, σ is the precision of the spatially resolved methane column density. The better the precision (i.e. smaller σ), the lower the detection limit.

The camera noise is determined in part by the amount of signal captured and detected by the instrument. For the case of passive (solar) illumination, the collected signal depends on the albedo a and the solar zenith angle θ_{sza}. Hence, the measurement precision in the ideal "shot-noise" limit scales as $\sim 1/\sqrt{a \cdot \cos(\theta_{sza})}$.

We must also account for the light path since it affects the total amount of methane "sampled" by the illuminating rays. This is done using the "air mass factor" $\mu \approx \frac{1}{\cos(\theta_{sza})} + 1$.

Combining these effects, we find that $Q_{50} \propto \sigma \propto 1/(\mu\sqrt{a \cdot \cos(\theta_{sza})})$.