

RC1

Ramier et al. present a detailed accounting of GHGSat's performance validation work during 2021-2026, specifically with respect to detection sensitivity and quantification accuracy. The validation assessment is based in single- and pseudo-blind controlled releases by the authors/GHGSat and third parties. The work largely leverages published methods but also presents unique insights into the relevant parameters affecting sensitivity in addition to a novel bias correction method for emission rate quantification.

I have some minor concerns about the underlying methodology of the paper. Specifically, the use of an idealized predictor; the limited assessment of alternative model formulations; and the extensibility of the results to other regions of observation (i.e., not the controlled release locations).

The paper is generally well-written, describes novel methods and results, and will be a relevant contribution to the readership of AMT. I recommend that it be accepted for publication after the comments below are addressed.

We thank the reviewer for their comments and recommendations. The revised manuscript contains an expanded list of models and further explains how the predictors were narrowed down. An explicit statement of scope limitations was added in the Conclusions and Discussion. Our responses to the individual comments follow below.

Main comments:

Abstract

Page 2, line 41 – can you say that the sensitivity of quantification accuracy is GLOBALLY insensitive to the wind product? Is this not specific to your controlled release locations? What about mountainous regions?

Indeed the intent was to say the quantification *metrics* derived from this study were not strongly affected by a change in wind speed source. This is now clarified in the abstract.

Additionally, the Conclusion and Discussion now explicitly mentions topography and land cover instead of the more vague “surface conditions”.

Methods

Page 4, line 93 – the authors claim that GHGSat is kept unaware of the “precise location” of controlled releases. To what resolution is this true? If the satellite has a spatial resolution of

27 m, how many pixels are truly viable candidates for a controlled release source location? I suggest that this language be softened if necessary.

The “precise location” statement was removed.

Page 5, line 120 – how does the anemometer height of 2.25 m compare with the wind speed product used for quantification. Is there some scaling required to bring this wind speed to 10 m before applying the quantification methods?

The effective wind speed used in the IME method indeed assumes a wind speed at 10 m above ground. The quantification section of the manuscript was updated and now includes figure entries for both the original anemometer wind, and a wind speed adjusted to 10m using a logarithmic wind profile for the CMC and Petrolia sites (other campaigns already measure the wind at 10m). Results using the height-adjusted wind are now used throughout (metrics of bias and scatter were updated), including in the bias correction section (Ueff coefficients also changed).

Page 7, line 144 – can the authors confirm whether plumes are identified manually or not and whether this aligns with standard GHGSat operating procedures? I.e., is this procedure relevant to GHGSat data products?

A statement was added that the emissions identification is done in accordance with GHGSat standard operating procedures.

Regarding the relevance to GHGSat products, the results from this article are applicable to “target mode” operations over known customer assets, as opposed to “survey mode” background monitoring of large areas. This aspect is discussed extensively in the Conclusions and Discussion section.

Detection performance

Page 9, eq 5 – I’m curious how false positives and true negatives affect these models? Wouldn’t a non-zero false positive rate imply that the POD does not go to zero at small q?

We added a paragraph at the end of section 3.1 which clarifies the interpretation of the POD as the conditional probability of detection, given the existence of a true emission. The paragraph discusses the relationship with other quantities of interest (marginal probability of a detection outcome, FP/TN conditional) which should address the reviewer’s questions. In short, FP and TN are described by a different term in the marginal expansion (which makes sense: the probability of a false positive should not be a function of the same variables such as emission rate or wind speed), but we don’t have enough data to analyze it beyond the number of outcomes.

“Predictor functions” – the chosen heuristic predictor (or detectability/SNR per Bruno et al.) is highly idealized. I understand the appeal of a dimensionless predictor, but should we expect this to fit the data best? Others using these methods seem to have found that the form of this predictor is sub-optimal. How does the lack of exponents on the variables in Eq. 6 affect the accuracy of the POD model? Have the authors tried alternative predictors that account for this?

We agree that the plain Jacob heuristic is idealized, and our results confirm this: it has the highest transition width, AICc, and BIC of all predictors tested, and the likelihood-ratio test now reported in Sect. 3.3 shows that adding a wind offset improves the fit significantly. The three alternative predictors in Table 2 are therefore already departures from the heuristic, but they are deliberately restricted to the wind dependence. We added a paragraph in Sect. 3.1.2 explaining why and putting this choice in the context of available literature.

I find the discussion on the retrieval noise (page 10) a bit confusing. Line 226 notes that the standard error of the retrieved methane nearby the source is used for n . The section then goes on to discuss column precision – where does this latter concept come in?

This section was moved to the new supplementary information section. We also edited the paragraphs that define the noise model, removed the word “precision”, which was used as a loose synonymous for “noise” or “uncertainty”, and make it explicit that $\hat{\mathbf{S}}_{x,r}$ is the relevant quantity for the rest of the analysis.

“Inverse link functions” – I’m surprised the authors only tested two inverse link functions (lognormal and log-logistic) as there are many other valid options used in the literature. Others (10.1016/j.rse.2023.113499; 10.1016/j.rse.2024.114435; 10.1021/acs.est.4c06702) have found other options fit better. Have the authors assessed other candidates and how they might improve their POD fit?

We expanded the list of inverse link functions to include the Fréchet and Weibull distributions. The results are listed in a modified Table 4, and in an additional table (Table S2) in Sect. S4 of the SI.

Page 14, line 320 – elaborate on the bootstrapping method; i.e., were the controlled release data resampled, then the same fits applied? This is a great inclusion, but theoretically only captures sampling uncertainties. Can the authors comment on the benefit (if any) of a Monte Carlo-sampling of the likelihood to characterize uncertainty?

We expanded the description of the bootstrap method, explaining the draw-with-replacement process and stating that we do a full maximum-likelihood fit on each of the sample subsets.

Regarding alternative approaches for uncertainty estimation, we agree that there are multiple approaches that could have been chosen. We chose the bootstrap method which is versatile and powerful despite its simplicity, and although it is based on sampling, it remains a valid approach to estimate uncertainty and confidence intervals. One approach we did try is error propagation based on the Hessian, which is a second order approximation to the likelihood function around the optimum and therefore tightly linked to the reviewer's suggestion. This yields a Q50 uncertainty of 5.2 kg/hr, which is very consistent with the 5.3 kg/hr obtained from bootstrapping. The Hessian-derived uncertainties are now listed in supplementary tables S2 and S3.

Quantification accuracy

Page 19, below figure 6 – shouldn't these comparisons be placed in the context of fit uncertainty? Is the modest improvement in the regression and reduction in scatter about parity when using measured winds statistically significant given overall uncertainty in the quantifications? This presumably has important implications for standard operations.

The language in the paragraph was softened, and the quantification improvements now contextualized relative to the high uncertainty of individual source rate estimates.

Bias correction – this is great, but I am curious how extensible it is outside of the controlled release sites and how sensitive it is to the procedure for segmenting the background from the plume.

We are actively working on a production-grade version of the bias correction method, which is robust to the wide variety of conditions and edge cases that can present themselves in more complex scenes. This is quite detailed and still under development, so will not be included in the paper to maintain the focus on controlled releases, but so far we find that the method generalizes quite well outside of the specific controlled release context. A note was added about this future work in the Conclusion and Discussion section.

Does the newly derived U_{eff} (Eq. 21 and page 21, line 470) use measured or modelled wind speed? If the former, how is it scaled from the local anemometer height to 10 m? Isn't this also highly specific to the controlled release sites? Would it be relevant in different terrains?

We use the local measured wind speed (now adjusted to 10m using a log profile), which is stated in the paragraph and labeled as the x axis of revised Fig. 6d, but the same method can be applied to any source of wind speed. We chose the local wind because the reduced scatter helps better visualize results and isolate the effect of the biases from the wind noise.

No height scaling was applied in the original submission, but the revised manuscript uses the 10m-adjusted wind for this analysis. The U_{eff} coefficients were changed accordingly.

The comment about the validity Ueff calibration is a good point, and we added it to the scope limitations listed in Conclusion and Discussion section near the end of the article.

Is it possible that your controlled release locations have relatively low biases and by applying this correction elsewhere one would artificially cause a high bias?

Those types of checks are part of the (already started) follow-up work mentioned earlier. So far, we do not see such failure modes – if anything, the more complex the scene, the more beneficial this method tends to be.

Page 21, line 480 – these corrections do rely on spatial resolution, the segmentation pipeline (i.e., what is background vs. what is plume), and the wind product (as noted on page 22, line 494), correct?

That sentence was rephrased to avoid implying the methods could be transferred as-is on other systems without any kind of tweaks.

Minor comments:

Page 2, line 28 – Ayasse et al. (2024; doi: 10.1021/acs.est.4c06702) evaluated the detection performance of EMIT.

We softened the abstract statement and scoped it to controlled-release-validated. We also now cite Ayasse et al. later in the article when describing the PoD model.

Page 3, line 52 – provide citation for the atmospheric lifetime of methane and the GWP.

The IPCC ref from the previous sentence continued to apply here (Lifetime: IPCC AR6 WG1 Table 6.2, GWP-20: IPCC AR6 WG1 Table 7.15). We now repeat it for clarity, and added a reference to WG3 (Mitigation of Climate Change) which is more appropriate to support the statement about methane being an important leverage. The citation key was also made more concise.

Page 3, line 64 – can the authors comment on the steady-state count of GHGSat-C satellites? This might be interesting and relevant to the reader.

The satellite count was updated to 16 (2 new instruments launched and 1 decommissioned during the review cycle). We prefer to withhold comments regarding the target “steady-state” number of instruments because of its dynamic, market-driven nature.

Page 6, figure 1 – I suggest including this figure in an SI. Are there similar figures of the other release sites?

This figure was moved to a new supplementary information document. We do not have pictures of the other sites to include, but such pictures are already publicly available in the cited references.

Bonus change: at first, this comment was mistakenly attributed to Fig. 2 of the pre-print (now Fig 1), so a new version of that figure was made with an example albedo and CH₄ for each of the 6 controlled release sites included in the study. We included a false negative in the mix (Fig. 1k) which is interesting to visualize next to a true positive (Fig 1l) that has a similar rate but less noise and lower wind. That figure was also moved a bit earlier in the manuscript (Sect. 2)

Page 7, eq. 1 – should this be a summation over pixels in the plume “mask” rather than an integral?

Equation (1) is now expressed as a summation.

Page 8, line 177 – there are many examples of measurement-based and -informed inventories in the literature that consider detection sensitivity. Consider citing some additional sources.

Four references were added.

Page 11, line 269 – I’m surprised that the Rice distribution did not work as well given its physical basis. I was hopeful; kudos for trying. Could the AICc of the various fits be provided in an SI (or added to Table 4) so that the reader could get a sense of how differently the various models performed?

A new table listing more detailed binary regression results was added to the SI, including some more additional link functions.

Page 12, line 282 – cite Conrad et al. (10.1016/j.rse.2023.113499).

The citation was added.

Page 15, line 325 – Q90 seems to vary much more than Q50 suggesting that the tails of the various POD fits can be quite different.

A comment about the larger variability of Q90 between models was added. We think this is more related to transition width than the tail of the distribution (Fig. 2a vs 2d) and there is a clear clustering between the plain heuristic and predictors that include a wind offset in one way or another (Fig. 3f). This effect is expected if the wind offset is correctly modeling a physical effect, as a better predictor model tends to produce a sharper transition. To illustrate this, consider the extreme cases: on one side, a very bad predictor will give random

detections and misses at any value of Z so the link function will be flat; at the other extreme, the predictor is fully deterministic and the link is a step function.

Page 16, line 358 – I’m very fond of this incremental assessment of each variable’s predictive power and the insights. Consider adding it to the abstract if space allows.

A sentence was added to the abstract highlighting this comparison and its outcome.

Page 17, figure 5 – there is an interesting spike in Q90 early in 2025. Can the authors comment on whether this is a new satellite performing very differently vs. additional controlled release data (I think the latter)?

This is simply the introduction of a new false negative sample (specifically on 2025-02-26, as part of the Stanford campaign). We reviewed it and found nothing particular about it that was worth mentioning in the article.

Page 18, line 402 – why is “true” in quotes? Considered replacing with metered as in the following paragraph.

The quotes were removed and this sentence lightly rephrased. The quotes were meant to reflect that even a flowmeter isn’t a perfect measurement and has some uncertainty.

Page 18, line 417 – clarify which specific wind data are used here and clarify in the caption of Figure 6 that the top and bottom rows are modelled and local wind speeds, respectively.

The wind models are now mentioned explicitly in the text and figure caption.

Page 21, line 457 – can “markedly” be quantified? E.g., the scatter at rates less than some kg/h reduce by some amount.

This paragraph was reworded to avoid redundant sentences, and the word “markedly” was removed as part of that change.

Page 23, line 511 – this is very interesting. Is this related to the skewness of the wind speed distribution?

Yes, the two effects are closely related. The skewness (Rice distribution or similar) arises from taking the magnitude of a random 2D vector, while the spatiotemporal dependence arises from the amount of averaging applied to such vectors before taking the norm.

RC2

This is a very interesting and important analysis about the best developed satellite methane system, and I enjoyed reading it. It is a comprehensive analysis that brings together

numerous studies. The manuscript has a solid basis in that it extends from previously published studies, but I feel that a number of communication-related issues would make it clearer and stronger:

We thank the reviewer for their comments and recommendations. We share the view that clarity supports the trust and engagement of the scientific community in this validation work, and hope the revised manuscript addresses the communication gaps. Our responses to the individual comments follow below.

Introduction: I suggest removing the final 6.5 lines of the Introduction because this information is repeated later in the Results and Discussion:

“Comparing multiple variants of this model based on the Akaike and Bayesian information criteria, we find the PoD to be well described by a log-normal function of the ratio of source rate (q) to retrieval noise (n), spatial resolution, and wind speed (u), with a linear offset parameter (u_0). The best-fit model obtained through maximum likelihood implies a 50% probability of detecting an emission rate of $99.0 \pm 5.3 \text{ kg h}^{-1}$ at a wind speed of 3 m s^{-1} and a median CH_4 column error of 7.0 mmol m^{-2} . We demonstrate that this estimate converges stably over time and remains highly consistent as independent blind validation is added to the self-organized releases. We also evaluate quantification accuracy through a parity analysis comparing GHGSat-estimated emission rates across all campaigns.”

The last paragraph of the introduction was reviewed, removing the numerical results which were redundant with the abstract and main text, and instead focusing on presenting the general structure of the article.

Table 1: The reference “Tarek 2026” is not included in the reference list.

This entry points to a unique reference, which should have read “FluxLab and Abichou, 2026” (first and last name were incorrectly parsed). It is now correctly listed in the text and bibliography.

Anon. reference: This protocol appears to be available online through the Carbon Management Canada website. If this is the same version, it could be properly referenced using the following web address: <https://cmcghg.com/wp-content/uploads/2025/09/MDAQ-Test-Protocol-Canada.pdf>

This reference is now listed as “CMC, 2025” in the main text and “CMC: Methane Detection and Quantification Testing Protocol Canada, Carbon Management Canada, <https://cmcghg.com/wp-content/uploads/2025/09/MDAQ-Test-Protocol-Canada.pdf>, 2025.” In the bibliography.

Section 3.1 is technically appropriate but very jargon-heavy for the broad community that may be interested in GHGSat outcomes. It is difficult to read without frequently stopping to think because the language is overly technical and some sentences are particularly challenging. For example: “The spatial resolution metric α is the 1-sigma width of the optical point spread function (PSF) in ground coordinates, scaled to account for pixel stretch at off-nadir viewing angles,” and “We note that the standard error based on the fit residuals provides a more accurate and conservative estimate of the local column precision than its ‘model noise’ counterpart which simply propagates sensor shot noise and readout noise into retrieval parameter space and therefore yields only a lower bound that can be exceeded due to correlated noise and poor model fit.” While I appreciate the economy of highly technical writing, some plain-language explanation would help here. These are not overly difficult concepts, and the authors could communicate them more clearly.

Section 3.1 was reworked to improve the clarity and flow. Jargon was dropped, softened, or complemented by a plain-language interpretation of the formal concepts.

Z is derived from a heuristic, but plume detection uses IME. Could the authors explain more clearly how a peak-based heuristic can be used to estimate PoD for an IME-based detection pipeline?

The IME is an essential piece of the *quantification* pipeline but is less closely related to plume detection. It describes the total amount of methane that raises above the noise floor, and therefore is undefined when there is no detection in the first place. The peak SNR heuristic, on the other hand, describes how visible the plume is relative to the noise floor, and thus closely connected to the likelihood of detecting it, either through automated or visual inspection.

Why is the false-positive side of detection performance omitted? As presented, the model characterizes only sensitivity rather than providing a complete performance specification.

The dataset does not have enough false positives ($n=1$) to conduct a rigorous analysis beyond stating the number of outcomes of each category (TP, FN, TN, FP), which are listed in Table 1. This limitation is now acknowledged explicitly in a new paragraph of section 3.1 (see also response to RC1 about Page 9, eq 5).

Section 3.2: Does the bootstrap method resample individual events, or is resampling blocked by campaign or site? Because the data are grouped by site and campaign, releases from the same site may share correlated noise and other characteristics.

The bootstrap method is now described more explicitly in the text. There was no grouping – samples are drawn from the full set of 125 TP/FN samples.

Figure 4 and surrounding text: Since resolution (a) does not improve the model on its own, please state explicitly that a in the equation is assumed rather than fitted.

The parameters of the PoD model are divided into 2 broad categories: environmental variables {q, u, n, a}, and model parameters { μ , σ , u_0 }. The model parameters are fitted, but all environmental variables (including a) are measured inputs and never fitted.

The differences among the predictor/link combinations are small, but the linear-offset model is adopted as the primary model. Perhaps a likelihood-ratio test against the Jacod heuristic model could be used to establish formal significance. How many parameters separate the two models?

A likelihood ratio test was added in section 3.3. The two models are separated by 1 degree of freedom: wind offset.

The text at the bottom of page 16, lines 364–373, could also be less jargon-heavy. The differences among the detection models were small and were smaller than the uncertainty caused by the limited set of experiments. Therefore, if the study were repeated using a somewhat different collection of methane releases, the model scores could appear in a different order.

The paragraph was rephrased to be more clear and less jargon-heavy.

Line 445: The range of 3 to 13 pixels is quite wide. Could the authors provide a sensitivity check?

A section of the new supplementary information document provides an analysis of the sensitivity to the background region parameters.

The manuscript frequently refers to a local anemometer. At the CMC site (Pseudo), the local onsite anemometer was 2.25 m above the ground, whereas wind measured at 10 m would be expected to be approximately 25% faster. Since Figure 6 presents results from Pseudo, Blind, and other campaigns, the authors must be referring to more than one local anemometer, presumably one for each study. At what heights were these other anemometers deployed? Were they all 2.25 m above the ground, or were some closer to the 10 m reference height of the global wind product? These are integrated into the modeling, so deployment characteristics and differences should be clear.

This is a good point and is also discussed in response to a comment by RC1. The quantification section of the manuscript was updated and now includes figure entries for both the original anemometer wind, and a wind speed adjusted to 10m using a logarithmic wind profile. The specific anemometer heights are now listed: CMC: 2.25 m, Petrolia: 7m, and all others directly measured the wind at 10m.

U_{eff} is presented in Figure 7. I am always interested by this adjustment factor. Based on the wind power law, a U_{eff} of 0.4 using an input wind of 10 m would correspond to a wind location well within the roughness layer. Realistically, many plumes are probably affected by terrain, vegetation, and other obstacles. However, one might expect U_{eff} to be closer to 1 in these relatively ideal controlled-release studies, after accounting for release height. In the CMC work, the vent stack heights were 3.4 and 4.8 m. Not very high and near the roughness layer. In the Stanford studies, I believe the release height was closer to 7 m. Do release height and proximity to the roughness layer affect U_{eff} and your model?

Those are very good questions and areas of active investigation, to which we can only partially answer. One important aspect is that the effective wind speed function used in the IME method is not simply the intuitive “average wind speed seen by the plume” it is commonly described as – it is an empirically calibrated scaling factor that also accounts for the plume length arbitrary definition ($L_{\text{eff}} = \sqrt{A}$) and specificities of the masking algorithm (e.g. threshold relative to noise floor). Because of this, we cannot make a direct correspondence between a certain U_{eff} factor (e.g. 0.4) and plume height above ground.

One could still expect relative height changes to matter (CMC 3.4m/4.8m VS Stanford 7m). However, quantifying how much it matters is not straightforward because the IME relies on the total visible mass of the plume, including enhancements downwind from the origin where methane has undergone significant vertical mixing (see for example Fig. 3b of Gorroño et al. 2026, 10.5194/amt-19-1245-2026). One kilometer downwind from the origin, a plume released from 3m or 7m are probably very similarly dispersed.

In short, the IME method is inherently empirical and makes the broad assumption that the conditions in which U_{eff} was calibrated (whether via simulations or controlled releases) are representative of field conditions. The method could definitely be improved to better account for specific sources, topographies or weather conditions, but this is left for future work.

The study develops a sensitivity outcome for these experiments, but the authors also develop a correction factor that extends these optimal results to less optimal real world conditions. The sensitivity is expected to be 1.6x worse in the real world. I think the implications in the real world should be clearer, not only in the abstract but in the conclusions as well. In the abstract, only the optimal value is listed. Many people read only the abstract, and may therefore get the wrong impression.

The abstract now includes the median Q50 evaluated in global conditions, and the Conclusions and Discussion section gained a dedicated paragraph about evaluating the Q50 in global conditions. A misleading explanation was also clarified: the quoted number

~160 kg/h is the median of Q50 evaluated over the ensemble of conditions, which is not exactly the same as the Q50 evaluated at median conditions as previously implied (this evaluates to ~150 kg/h).