Response to the Referees' Comments

REFEREE #2

Half the metrics are directly related to detection (TP, FP, FN, TN, N(L=5), Lbar), however how is detection actually done in the different methods?

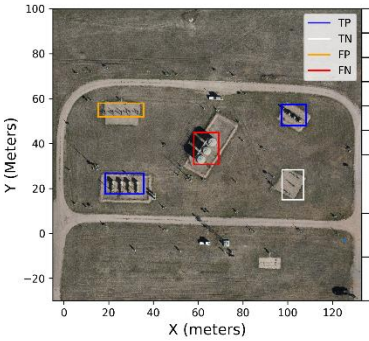
Authors:

Thanks for your comment. We have added a case study experiment to illustrate an example experiment, associated release rates, estimates, and evaluation to help clarify the meaning of some of these metrics.

Changes:

To aid in the understanding of these metrics, an illustrative example of these evaluative metrics applied to a single experiment from the study is shown in Figure 3. This figure shows an image and two tables that summarize the output of the system (rate estimates for each equipment group) alongside the ground-truth release rates (top table) and computes the relevant evaluative metrics for this single experiment (bottom table). During this experiment, there were three active release sources: the tanks (group 4T), the western separators (group 4S), and the eastern separators (group 5S). The western and eastern wellheads (groups 4W and 5W, respectively) were not emitting. The quantification estimates are shown in the 2nd column of the top table, while the ground-truth release rates are shown in the 3rd column of the top table. The final column shows the classification of the estimate as either a TP/FP/FN/FP as previously described. The table on the bottom shows the relevant evaluative metrics applied to the estimated and ground-truth rates in the top table. We see that, for this example experiment, there were 2 true positives (the system accurately identified that both the 4S and the 5S groups were emitting), one false negative (the system missed that the tanks were emitting), one false positive (the system assigned a small but nonzero rate to the 4W group, which was not emitting), and one true negative (the system accurately identified that the 5W group was not emitting). These statistics are summarized in the bottom table, along with the overall "localization score", which in the case, was 3 (i.e., the emission status of 3 out of the 5 equipment groups were correctly identified). The total estimated and actual facility-level emission rate is shown in the bottom table as Q and Q' (these are computed as the sum down the "Estimated Rate" and "Actual Rate" columns, respectively). In this example, the estimated facility rate is 1.73 kg/hr while the actual emission rate is 1.83, representing an error of -0.1 kg/hr (E) and a relative error of -0.055 (i.e., -5.5% error, E_{rel}). In terms of the other quantification-related metrics ($F2$ and ΔC), this experiment's estimated facility-level rate is within a factor of 2 of the actual rate (so it would positively contribute to the fraction of estimates that were within this factor, when summing over all experiments), and the contribution to the cumulative error from this experiment would simply be $E.\Delta t$, where Δt is

the duration of this experiment. The duration of this particular experiment is 30 minutes, so the contribution to ΔC is -0.05 kg.



Group	Estimated Rate	Actual Rate	Classification
4S	0.26	0.52	TP
4T	0.0	0.81	FN
4W	0.05	0.0	FP
5S	1.42	0.5	TP
5W	0.0	0.0	TN

TP	FP	TN	FN	L	Q	Q'	E	E_{rel}
2.0	1.0	1.0	1.0	3.0	1.73	1.83	-0.1	-0.055

Figure 3. Example experiment to illustrate the evaluation of the output of the system with respect to ground truth rates. The image on the left shows each equipment group's estimate classified as either a TP/FP/FN/FP. The upper table summarizes the estimated rates, actual rates, and the detection classification, while the lower table applies the evaluative metrics described above to the data from the upper table.

A lot of space spent on the models and metrics, comprising Section 3, a lot of which is already described in the literature. I think it could be shortened, if possible, to better highlight the results Section 4. A key novelty of the manuscript is the multi-source estimation estimate, along with large number of experiments with continuous monitors. Many analyses could be envisioned, in particular detection curve vs emission rate, whether any equipment groups perform better (perhaps due to prevailing wind patterns or other factors), simulating if there were fewer sensors (as mentioned might be realistic), interference (if small leaks sometimes are hidden by larger ones), effect of experiment time vs DLQ accuracy (30 minutes vs 8 hours), etc.

Authors:

We moved parts of the Introduction and Methodology sections to the Appendix to make the paper more concise.

Changes:

Please see Appendices A and B.

L43 add AVO in parentheses

Authors:

We added the abbreviation in parentheses.

Changes:

Traditional approaches for detecting methane emissions often rely on human senses (auditory, visual, and olfactory (AVO) inspections) or portable sensors used in close proximity to potential sources.

Cheptonui 2024 a/b are same paper

Authors:

We corrected the reference.

Changes:

Several studies have independently evaluated the efficacy of CMS in quantifying emissions, suggesting promising advancements in recent years (Bell et al. 2023; Ilonze et al. 2024; Cheptonui et al. 2025).

More details on the data collected as part of the 2024 CSU METEC controlled release study can be found elsewhere (Cheptonui et al. 2025).

Sec 2 – for releases at multiple equipment groups, does each release rate simply randomly belong to the overall distribution in Fig 1b?

Authors:

That is correct. We have added a figure and some explanation that more concretely shows the releases from a single experiment, the output of the system, and how this particular case is "scored".

Changes:

The experiments are designed such that only one release point is active per equipment group at the METEC facility. Each equipment group is composed of numerous “equipment units” (i.e., individual tanks, wellheads, or separators) and each equipment unit may have multiple potential release points on it. In other words, each equipment group has numerous \textit{potential} release points, but only one is ever active at a time for a given experiment. In this study, we focus on the ability of the system to correctly detect, localize, and quantify to the equipment group level. As such, the centroid of each equipment group is computed and these 5 coordinate pairs (corresponding to the 5 equipment groups at the facility) are used as the potential source locations as an input to the localization and quantification (LQ) algorithms.

L270 I am surprised to characterize the importance of the stability class and dispersion coefficient parameterization as minimal. Sure they may be representing the same behavior, but don’t they empirically have a significant effect?

Authors:

To clarify, the stability class and dispersion coefficients indeed play a large role in predicting concentrations (and hence, quantified flux rates). However, many of the parametrizations of them are simply different empirical formulae, derived from the same underlying publicly available data, and give very similar approximations of the stability class and dispersion as a function of distance. If a poor parametrization is used, then the performance of the quantification will be poor. However, most of the commonly accepted and standard methods are very similar to one another and are just different functional forms and associated coefficients. We have added a sentence clarifying this statement

Changes:

The following subsections provide brief overviews of the theory underpinning the dispersion models, followed by more specific implementation details. Note that there are myriad small choices (e.g., stability class calculations, dispersion parametrization) that must be made in the data processing and algorithmic workflow when it comes to running these dispersion models. It is outside the scope of this study to enumerate and present results from every combination of valid choices. Instead, we will provide clear justifications for the specific choices made in this study and demonstrate the efficacy of the models under these specific implementations. It should be noted that the impact of most of these higher-order decisions on the results is minimal, as they are often different approaches of approximating the same underlying phenomena. For example, there are several commonly-used functional forms and associated coefficients to describe how the dispersion of a gas plume scales with distance. While these empirical formulae may look very different (e.g., some utilize power laws while others employ logarithms), they are generally inferred by fitting these functional forms to the same underlying data, and result in similar general characteristics despite the sometimes dramatically different functional forms.

L275 GMP -> GPM

Authors:

Thanks for your note. We corrected the error.

Changes:

The commonly used Gaussian Plume model (GPM) provides a closed-form solution to the steady-state advection-diffusion equation for a single point source emitting at rate Q from height H.

Eq 8 (for the Gaussian puff) seems unusual – please check. Normally there is a 2*pi^(3/2) factor, and importantly, time dependence

Authors:

Thanks for your comment. The correct exponent of 3/2 on the 2*pi has been added to the equation, and the explicit time dependence is mentioned.

Changes:

$$C(x,y,z,t) = \frac{Q\Delta t}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z} \exp\left[-\frac{y^2}{2\sigma_y^2} - \frac{(z-H)^2}{2\sigma_z^2} - \frac{(z+H)^2}{2\sigma_z^2} - \frac{x^2}{2\sigma_x^2}\right]. \quad (8)$$

L439 Is a no flux condition typical in these simulations? A 200 m boundary seems like it would significantly affect the dispersion behavior

Authors:

Thanks for your note. If we were considering much larger length scales, then the 200m zero-flux upper boundary may artificially impact on the results, however here, the average source-sensor distance is around 50 meters, and the releases are effectively at ground level. As such, virtually no simulated gas reaches the "ceiling" of the simulation within the relevant length scales of source->sensor.

Changes:

No changes were made.

Fig 5 some information is not visible on this plot (puff LSQ on L=4 and plume MCMC on L = 1)

Authors:

Thanks for your note. The plot has been modified to include variable line widths and marker styles so that overlapping lines are more distinguishable.

Changes:

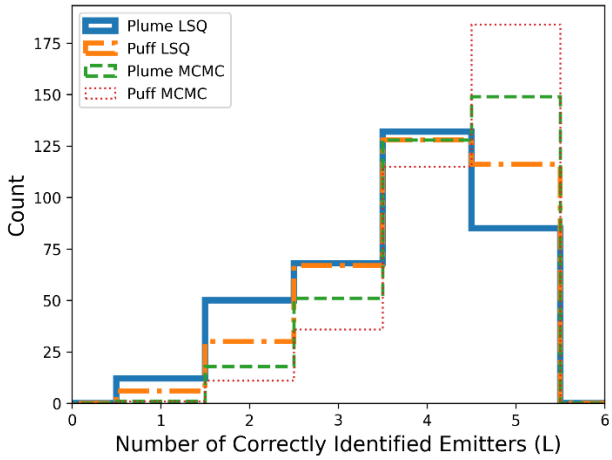
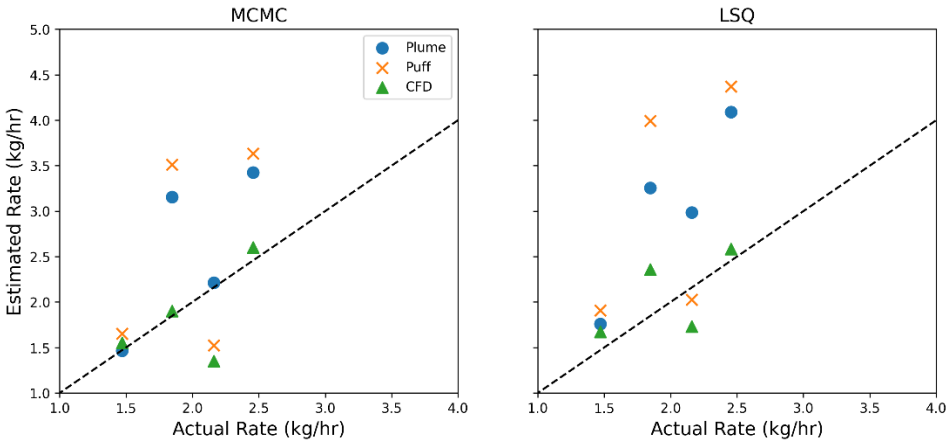


Figure 11 suggest using different marker types, as the colors by themselves can be difficult to distinguish

Authors:

Thanks for your note. Each model now has a unique marker in the plot and the marker size has been increased to help with visibility.

Changes:



L173 Inverse distance weighting is mentioned, and "sonic anemometers" (plural) here, but is unclear to me where/how many where used. Could this be added to Fig 2 or otherwise?

Authors:

Thanks for your note. We have added a sentence clarifying that each of the sensors in this study was equipped with an anemometer in Section 2 and also added some wording around this in discussing the inverse distance weighting.

Changes:

Ten Canary X integrated devices were installed within the METEC site perimeter to measure ambient methane concentrations. All of the Canary X devices used for this study were equipped with an anemometer.