Reply on first review

Machine learning-based emission rate estimates of global methane super-emissions

Roberts et al.

Please address all correspondence to Clayton Roberts (*c.roberts@sron.nl*)

Summary

We thank the two reviewers sincerely for their time and expertise in reviewing the first version of the manuscript. The revised manuscript is better as a result of the reviewers' comments and suggestions.

Reviewer R1 requires some clarification to represent cited work more accurately, and an increased discussion on the behavior of the IME source rate quantification and its sources of uncertainty, and recommends that the manuscript be published subject to minor revisions.

Reviewer R2 asks that some uncertainties and potentials for future improvement be stated more clearly in the abstract and conclusions, and more specifically for improvements to discussions on the Karaturun case study and area sources, and recommends that the manuscript be published subject to minor revisions.

We have responded to reviewer suggestions in the most part by modifying the text. Where we feel the original text is more suitable, this rejoinder provides our reasoning. We have also thanked the reviewers in the acknowledgements.

In the following sections, reviewer comments are given verbatim in italics and referenced by reviewer and comment number (prefixed 'R' for first review, and 'RR' for second review, etc.). Our replies (prefixed 'C', 'CC', etc.) are given in normal font per review comment. We reference new locations in the article using line numbers from the latest version of the manuscript that incorporates reviewer comments (without tracked changes), to avoid confusion with page- and line-number differences between manuscripts.

Reviewer 1

R1.1 Line 1, the abstract states methane to be the "second most important greenhouse gas." This neglects the inclusion of water vapor as a greenhouse gas. Suggest amending to "second most important anthropogenic greenhouse gas."

C1.1 We agree with your point and have amended the first sentence of the abstract to read as "Methane, the second most important anthropogenic greenhouse gas, has a global warming potential more than 80 times that of carbon dioxide over a 20-year period." (Lines 1-2)

R1.2 Lines 71-73, Bruno et al. uses 10 m wind speed as a CNN input and is therefore not applicable to this set of citations for eliminating dependence on meteorological datasets.

C1.2 We apologize for the error and have removed the reference to Bruno et al. (2024) from this sentence. The sentence now reads "In the context of methane remote sensing, CNNs have recently been applied to estimate point source emission rates directly from plume imagery (in addition to detection applications), offering the potential to reduce or eliminate dependence on meteorological datasets (Jongaramrungruang et al., 2022; Plewa et al., 2025)." (Lines 76-78)

R1.3 Lines 76-77, Plewa et al, Joyce et al, and Bruno et al, also rely on simulated plumes for training. It is unclear why Jongaramrungruang et al and Radman et al. are chosen as the specific citations for this claim.

C1.3 We apologize for the confusion and have also included these additional citations at this location. The sentence now reads "Because real satellite observations never come with precise known emissions (except for controlled release experiments), training typically relies on simulated plumes (Jongaramrungruang et al., 2022; Joyce et al., 2023; Radman et al., 2023; Bruno et al., 2024; Plewa et al., 2025), where the emission rate is known for each generated image." (Lines 80-83)

R1.4 Lines 116-117, Some justification for the chosen injection heights for simulation of 25 and 250 m would improve this section. 25 m is reasonable for something like a flare stack, but is difficult to imagine what infrastructure would be injecting at 250 m. If the two heights are simply chosen as sensitivity testing to the injection height that is fine but should be stated.

C1.4 We have added a sentence at lines 125-126 clarifying this choice, which reads as "We choose these release heights to mitigate potential model overfitting to a single release height within the planetary boundary layer."

R1.5 Lines 141-147, echoing the earlier comment, Bruno et al. used a multi-channel input including a plume mask and wind channel. Language should be adjusted to reflect the fact that not all of the high-resolution CNN approaches relied on plume morphology for wind information.

C1.5 We apologize for the confusion and have included a sentence at line 151 that reads as "Some high-resolution CNN approaches still explicitly incorporate wind information (Bruno et al., 2024)."

R1.6 Lines 202-204, Why is uncertainty on the TROPOMI retrieval not included in this analysis?

C1.6 In our uncertainty analysis, we assume that the spatially varying methane background within a TROPOMI scene is fully correlated. This allows us to represent the uncertainty in the scene background with a single value, defined as the standard deviation of non-plume pixel values. The distribution of these non-plume pixels inherently captures the TROPOMI retrieval error. Moreover, assuming a fully spatially correlated background is a more conservative approach (i.e., it yields larger uncertainty) than alternatives such as using the standard error of the mean of the background pixels or propagating the TROPOMI retrieval uncertainties of the background pixels to estimate the uncertainty in their mean. We have added a sentence explicitly stating this at lines 224-227 which reads as "We adopt this background reduction approach because it is more conservative (i.e., it yields a larger estimate of background uncertainty) than defining background uncertainty as the standard error of the mean of background pixels or propagating the TROPOMI retrieval uncertainties on those pixels to the background mean."

R1.7 Lines 235-251, How you define Ueff is extremely important to how well IME will do on plume quantification. This section would be greatly improved by showing that the Ueff from Shuit et al. is a match for the fit calculated from this study's training dataset. Expanding on this, Ueff fits are often poor in the <3 m s-1 wind range, where this study finds marked improvement over IME for quantification (see Varon et al. for example where a log fit instead of linear fit is used to attempt to account for this). Showing how the Ueff calculation fits the data from rearrangement of eq (1) would help clarify the source of poor performance of IME in these low wind speed scenarios.

C1.7 We thank the reviewer for emphasizing the importance of the effective wind speed (Ueff) parameterization in IME-based plume quantification, particularly at low wind speeds.

We agree that Ueff strongly influences IME performance. However, improving the IME method or recalibrating the effective wind speed is beyond the scope of this work. Our objective is to develop and benchmark ML-SPERE against a published IME implementation that has been specifically calibrated for TROPOMI observations. Accordingly, we adopt the IME method and Ueff parameterization from Schuit et al. (2023b), which is also used operationally to estimate emission rates for TROPOMI methane plumes distributed to CAMS (Copernicus Atmospheric Monitoring Service, 2025).

We acknowledge that Ueff parameterizations can perform poorly at low wind speeds, as noted in prior studies, and that this likely contributes to the reduced IME performance observed here. However, revising the Ueff formulation is beyond the scope of this study, which focuses on method comparison rather than IME development.

We have added the following sentences at lines 264-267 to summarize this approach: "We apply the IME method and U_{eff} parameterization from Schuit et al. (2023b) without modification, allowing us to benchmark the performance of ML-SPERE against a published and operationally used implementation of the IME method that was specifically calibrated for TROPOMI methane observations."

R1.8 Lines 297-298, Language here needs to be cleaned up/clarified. The biasing towards mean emissions is indeed a common feature of CNN based source rate estimators. However, the assertion "A similar (and stronger) trend is observed for IME estimates. This tendency for predictions to regress toward the mean is a common issue reported in similar studies (Jongaramrungruang et al., 2022; Joyce et al., 2023; Bruno et al., 2024)," requires more justification/discussion. Previous work does not suggest IME to be a mean biased predictor. In fact, it is difficult to intuitively understand what such a behavior would refer to. Mean biased towards the mean of the dataset used to calculate Ueff perhaps? Please clarify if the claim is that IME has a stronger mean biased behavior than the new CNN approach, and if so justify how this might be the case as IME is not "trained" in the same sense as the CNN.

C1.8 We apologize for the confusion and did not mean to imply that the IME method is a mean-biased method. We only intended to point out that, despite not being a mean-biased method, the IME method implemented in Schuit et al. (2023b) (and used in this work) also shows a negative bias in emission rate estimates (for plumes with large emission rates) when applied to our test set. We have altered the text at lines 321-325 to read as follows: "The IME method also tends to underestimate emissions for plumes with large emission rates, despite the fact that the IME method is not a mean-biased method. The tendency for ML-based predictions to regress toward the mean of the training dataset labels is a common issue reported in similar studies (Jongaramrungruang et al., 2022; Joyce et al., 2023; Bruno et al., 2024). This regression dilution for ML-SPERE is reduced by 1/3 when we restrict application to well-observed plumes (Fig. A4**d**), though no such improvement in estimate biases is found for the IME method (Fig. A4**c**)."

R1.9 Lines 340-341, The statement, "For the four dates in the time series where we have an IME emission rate estimate from a high-spatial resolution satellite, we find greater agreement with ML-SPERE estimates than we do with TROPOMI IME estimates." Appears to only be true for 3/4 dates as the Jul 11 PRISMA emission rate is closer to the IME estimate than the ML-SPHEER estimate. Text should be amended to reflect this.

C1.9 We have altered the sentence on lines 366-368 to read as "For three of the four dates in the time series where we have an IME emission rate estimate from a high-spatial resolution satellite, we find greater agreement with ML-SPERE estimates than we do with TROPOMI IME estimates."

R1.10 Lines 446-448, the statement "model. We show empirically that this is because ML-SPERE has learned to use methane gradients around the plume head to inform emission rate estimates, which lends ML-SPERE clear physical interpretability," is the first mention of gradients around the plume head in the main text of the article. This is addressed more completely in the supplement but must be addressed in the preceding text to justify making this claim in the conclusion.

C1.10 We discuss earlier in the text that ML-SPERE obtains much greater performance than the IME method for plumes where we have complete spatial information around the plume head. We have altered the text at lines 296-298 to now explicitly mention the methane gradient at the plume head, and the text now reads as "Additional figures and analyses in Sect. A3 show that the improved performance observed when restricting the application of ML-SPERE to such plumes arises from the model's strong reliance on spatial gradients in methane abundance near the plume head or source region when inferring emission rates."

Reviewer 2

R2.1 Lines 8f: I stumbled on "the bias" because it's the first time the bias is mentioned. I suggest to rearrange the sentence, e.g. "IME exhibits a bias at low wind speed; ML-SPERE does not"

C2.1 Thank you for this suggestion; we have altered the sentence at lines 8-9 to read as "At low wind speeds, IME estimates exhibit a negative bias and ML-SPERE estimates do not."

R2.2 Line 25: To round out the argumentation, you could insert a sentence on the relevance of observations for mitigating super-emitters.

C2.2 Thank you for this suggestion; we have added a sentence (and citations) at lines 26-28 to express this point, which reads as "Satellite observations have thus proven instrumental in detecting and abating methane emissions from super-emitters (Irakulis-Loitxate et al., 2022; United Nations Environment Programme, 2025)."

R2.3 Lines 33 and 191: Since you bring up computational cost as one of the benefits of machine learning-based methods in the introduction: Can you add a few details about the computational cost of ML-SPERE vs IME? Line 191 has a bit on the cost of ML-SPERE training, but there is nothing concrete about the cost of IME in the text.

C2.3 We have added four sentences on lines 285-289 which read as "In terms of computational efficiency, the IME method implementation from Schuit et al. (2023b) is faster than ML-SPERE. On a standard desktop machine, it takes approximately 15 seconds to generate either an ensemble estimate of 43,923 members with the IME method or an ensemble estimate of 5,000 members with ML-SPERE. These timings may vary depending on hardware. The time it takes ML-SPERE to produce an ensemble estimate scales linearly with the number of ensemble members."

R2.4 Lines 51f: "converting satellite-retrieved methane columns into plume enhancements and accurately distinguishing plume pixels from background concentration": This sounds like two different sources of uncertainty and I think you mean "by accurately distinguishing", which would be easier to read.

C2.4 We apologize for the confusion and have altered this sentence to make it clearer that these are separate sources of potential uncertainty. This sentence on lines 54-56 now reads as "Potential sources of uncertainty include the conversion of satellite-retrieved methane columns into plume enhancements and the subsequent delineation of plume pixels from background pixels (Varon et al., 2018; Schuit et al., 2023b)."

R2.5 Lines 64f: "If the plume can be accurately modeled, inversions in general provide the most accurate emission rate estimates". That's a big "if" and therefore, I'm not sure the statement is both relevant and correct. Can you provide a source where the performance of mass-

balance methods vs. inversions was systematically assessed? In general, inversions, just like mass-balance methods, rely on the accuracy of modeled winds, plume rise, etc. Specifically, the common discrepancy between modeled and real wind fields can lead to displacements of modeled plumes, leading to a double-penalty for modeled emissions and thus underestimation and/or smearing out of point-sources by inversions. Mass-balance methods don't have this particular issue. Unless you can corroborate the statement, I would remove it.

C2.5 Thank you for raising this point. We have altered the text at lines 67-70 to read as "Compared with mass-balance approaches, Bayesian inversion methods provide emission-rate estimates that are more explicitly constrained by atmospheric transport physics and are therefore generally better suited to complex meteorological conditions and source configurations (e.g., overlapping sources), albeit at substantially greater computational cost and under the assumption that atmospheric transport can be modeled accurately."

R2.6 Lines 81f: "This is due to the fact that": not entirely clear what "this" refers to. Perhaps "The method relies on the fact that" or something like that.

C2.6 We apologize for the confusion and have altered the text at lines 87-89 to read as "Such CNNs can obtain comparable (or superior) performance to the IME method without using wind data because high-resolution imagery can resolve fine-scale turbulent structures within methane plumes, which implicitly encodes wind speed and direction (Jongaramrungruang et al., 2022; Joyce et al., 2023)."

R2.7 Line 98: Suggestion: "previously analyzed using inverse modeling" [change to] "previously analyzed using inverse modeling and compared to IME estimates based on GHGSat data"

C2.7 Thank you for your suggestion. We have included it with a slight alteration to reflect that Guanter et al. (2024) included IME analysis from other high-resolution satellites apart from GHGSat; the sentences on lines 103-106 now read as "Additionally, we contrast ML-SPERE and IME emission rate estimates for a TROPOMI plume dataset previously analyzed using both inverse modeling as well as IME estimates based on high-resolution satellite observations (Sect. 3.2; Guanter et al. (2024)). We also contrast ML-SPERE and IME emission rates estimates for all 2021 TROPOMI super-emitter detections from Schuit et al. (2023a) (Sect. 3.3)."

R2.8 Lines 115f: While you claim that your training and validation dataset covers "varied surface and meteorological conditions", it strongly focuses on northern hemispheric summer (6/7 simulated domains), and specifically, on areas close to the Caspian Sea (5/7 domains). I'm concerned that this choice might lead to overfitting (performing better on scenes that are close to the conditions at the training scenes and worse compared to others). Appendices A5 and A6 partly address this concern and you conclude that training set diversity is much less important than plume morphology for the performance. However, the argument is based on cross-validation within the limited training dataset, so I'm not fully convinced that geographic and seasonal limitations in the training data play only a negligible role in the performance. One could further examine the performance of the model on varied scenes. However, the performance advantage of ML-SPERE vs. IME is clearly shown in Sect. 3.1 and results are similar on average as for IME for a large variety of scenes (Fig. 6). Therefore, I don't think it's necessary to dissect the performance of the model further in this paper. But I suggest that you consider including an outlook with potential improvements in the conclusions, and that one of them may be a more varied training/validation dataset (i.e., other domains and seasons). For example, a future training set could cover situations where the current ML-SPERE and IME differ (the ones in Sect. A8). More on that outlook in my comment on lines 429ff.

C2.8 Thank you for this suggestion. We have an included a sentence in the conclusions at lines 495-498 which reads as "Future work could examine re-training ML-SPERE using a larger dataset, with expanded geographic and seasonal diversity that focuses on regions and circumstances where we have shown that ML-SPERE and the IME method produce larger differences in emission rate estimates, though the results of Sect. A5 show that ML-SPERE generalizes beyond the spatial domain of training data."

R2.9 Lines 122f: The split of the simulated plumes into training and validation dataset was random (ratio 90/10). Upon first reading, I was concerned that including the same domains in both training and validation could lead to overfitting. Please add a reference to Section A5 here, where this concern is addressed and quantified.

C2.9 We have included a reference to Sect. A5 at line 131 to address this: "We show in Sect. A5 that this choice of split is resilient to geographic overfitting by ML-SPERE."

R2.10 Line 131: Would it not be better to use a uniform distribution, to alleviate the regression dilution you observed? Fig. 7g of Schuit et al. suggests that the high emitters are severely underrepresented if the training dataset follows that distribution. If you agree, consider including that as an outlook.

C2.10 This is a valid question, though we believe our choice is well-motivated and we explain our reasoning below.

We chose the distribution of our training labels (i.e., the emission rates of the simulated plumes in our training dataset) to closely match the distribution shown in Fig. 7g of Schuit et al. (2023b). This means that during model training, ML-SPERE will be less often exposed to plumes with very high emission rates (e.g. greater than 100 t / hr) than at lower emission rates (e.g. in the range of 10-40 t / hr). This means that ML-SPERE is specializing (slightly) in the emission rate range where we expect to actually observe super-emitters with TROPOMI most frequently; in a sense, we are incorporating our domain expertise into our model training with this choice of training label distribution.

If we were to change our training label distribution to a uniform distribution (say in the range 0 - 200 t / hr), this would shift the neutral point of the regression dilution to 100 t / hr; i.e., a plume with an emission rate of less than 100 t / hr may potentially be overestimated by ML-SPERE, and a plume with an emission rate greater than 100 t / hr may potentially be underestimated by ML-SPERE. In the absence of any domain knowledge about the true distribution of super-emitter emission rates globally that TROPOMI can detect using the methods of Schuit et al. (2023b), this would be a fine design choice, however the results of Schuit et al. (2023b) inform us that in reality, this would likely mean that most plumes passed to ML-SPERE may have their emission rates overestimated. The distribution shown in Fig. 7g of Schuit et al. (2023b) is of course only an estimate of the true emission rates of the superemitters detected by the pipeline presented in that work, however we take that distribution as sufficient evidence to motivate the choice of our training label distribution. This is further motivated by the fact that we apply ML-SPERE to plumes which are detected by the pipeline presented in Schuit et al. (2023b), as opposed to an alternative TROPOMI methane plume detection scheme.

R2.11 Lines 134f: "The pixel containing the plume source remains within the central 28×28 pixel region of the cropped scene" - Does that mean that in some cases, the plume is only a few pixels long because it is close to the downwind edge of the scene?

C2.11 Yes, this means that some plumes could be only a few pixels long because they are close to the downwind edge of the scene. More often than not however, the smallest plumes (i.e. those only 2-3 pixels in area) are plumes with weak enhancements relative to the scene background variability. The choice of exposing ML-SPERE to plume head locations within the central 28x28 region was made to ensure that ML-SPERE does not require plumes to be centered within a scene before being passed as input. We have altered the sentence on lines 142-144 to state this explicitly: "The pixel containing the plume source remains within the central 28×28 pixel region of the cropped scene, which ensures that ML-SPERE does not require plumes to be centered within input scenes after training is complete." On lines 284-285 we further clarify that "We find no significant change in bias for either method when applied only to plumes whose masks are truncated by the scene boundary."

R2.12 Lines 142f: I'm amazed that the attempt without wind data lead to anything but noise at all. Please specify the statement "30% lower relative performance than models incorporating wind channels" - I assume "lower relative performance w.r.t. the test dataset"? Out of curiosity: Do you know how the model learned to distinguish wind speeds in the absence of wind channels? I guess it must have something to do with plume morphology after all, though I didn't expect that there is that much information on that at the scale of Tropomi pixels.

C2.12 We apologize for the confusion and have altered the text at lines 152-153 to specify that the relative performance decrease is with respect to final model performance: "Our initial attempts to train a CNN for usage with TROPOMI observations without wind information exhibited over 30% lower relative performance on our test dataset than the final model which incorporates wind channels."

The fact that these early models exhibited a weak ability to estimate emission rates from TROPOMI plumes without wind data indicates that there is some wind information from plume observations, though not enough to motivate excluding wind data from model training. It is possible that the model is learning that larger / longer plumes often correlate with higher emission rates, as it is difficult to delineate plume pixels from background pixels at lower enhancements.

R2.13 Lines 166ff: Please specify how many unique backgrounds were used. E.g. one per scene (224 unique backgrounds) or one per plume (438 unique backgrounds)?

C2.13 We have altered the text at lines 176-178 to reflect that a unique background scene was used for each test set plume: "To enhance the representativeness of the test set scenes and capture the challenges involved in estimating methane enhancements in TROPOMI data, we add a unique 32×32 pixel TROPOMI observation as background methane to each scene (Jongaramrungruang et al., 2022; Bruno et al., 2024)."

R2.14 Line 214: A comment (no need to change anything in the manuscript): I think the background uncertainty could be overestimated by sampling from the standard deviation and not the standard error of the background pixels (sigma/sqrt(n)). Though the number of pixels n would be debatable.

C2.14 We agree, and have deliberately chosen this approach in order to retain a conservative (i.e. larger) estimate of uncertainty. In our uncertainty analysis, we assume that the spatially varying methane background within a TROPOMI scene is fully correlated. This allows us to represent the uncertainty in the scene background with a single value, defined as the standard deviation of non-plume pixel values. The distribution of these non-plume pixels inherently captures the TROPOMI retrieval error. Moreover, assuming a fully spatially correlated background is a more conservative approach (i.e., it yields larger uncertainty) than alternatives such as using the standard error of the mean of the background pixels or propagating the TROPOMI retrieval uncertainties of the background pixels to estimate the uncertainty in their mean. We have added a sentence explicitly stating this at lines 224-227 which reads as "We adopt this background reduction approach because it is more conservative (i.e., it yields a larger estimate of background uncertainty) than defining background uncertainty as the standard error of the mean of background pixels or propagating the TROPOMI retrieval uncertainties on those pixels to the background mean."

R2.15 Lines 216-222: A comment (no need to change anything in the manuscript): I'm concerned that this method could lead to unphysical wind fields and unphysical combinations of wind field and plume morphology. However, since this method is "only" applied to obtain an uncertainty estimate, I would leave it as is.

C2.15 Our usage of a wind magnitude scaling factor drawn from a truncated normal distribution (instead of a "normal" normal distribution) should ensure that the resulting wind vector field should always remain physically realistic, however your point regarding the combination of the resulting wind field and methane scene is valid; we purely do this to estimate uncertainty.

R2.16 Line 265: By "constant absolute error", I think you mean a "constant bias"? Please reference the plot that shows it (Fig. 3d) in the text. Also, the linear regression line shown in Fig. 3d doesn't capture the "constant" bias described in the text, of course. To clarify, could you add density maps to panels c and d, similar as panels a and b? It's a detail, but it could help to match the description in the text to what is shown in the figure.

C2.16 We apologize for the confusion; the absolute error we are referring to is not the trendline in bias shown in Fig. 3**d** in the text. What we intended to state is that the absolute errors produced by ML-SPERE tend to scale with the emission rate of the plume, however, this trend does not hold as we scale down to plumes with smaller and smaller emission rates. Below a threshold of approximately 20 $\mathrm{t\ /\ hr}$, the average absolute error asymptotes to 7 $\mathrm{t\ /\ hr}$. This means that in percentage form, absolute errors at lower emission rates become inflated. Fig. 3**d** is showing the average linear trendline fitted through all ML-SPERE residuals for the entire test set, which does not capture this behavior. We have altered the text at lines 280-282 to simplify and clarify this point: "Absolute errors produced by ML-SPERE generally scale with the true emission rate of test set plumes, but this relationship breaks down at the lowest emission rates, where an approximately constant absolute error leads to inflated relative errors. Similar behavior is observed for IME estimates."

R2.17 Lines 270-282: Here and in Table 1, please also describe the performance of ML-SPERE vs IME for plumes that are not well observed.

C2.17 We have included new text on lines 303-306 to describe the performance of ML-SPERE and the IME on only the test set plumes which are not defined as being well-observed: "For poorly-observed plumes in the test set (i.e., those that do not meet the criteria for being well-observed), ML-SPERE and the IME method exhibit broadly comparable performance. ML-SPERE performance degrades relative to its results on well-observed plumes, whereas IME performance remains similar across all subsets." We have also included summary statistics reflecting this in Table 1 on page 11.

R2.18 Line 293 and Line 584: However, the tests in A5 and A6 do not examine the limited seasonal coverage of the training/validation dataset. See my comment on lines 115f.

C2.18 We have included an outlook statement at lines 495-498 to reflect this point: "Future work could examine re-training ML-SPERE using a larger dataset, with expanded geographic and seasonal diversity that focuses on regions and circumstances where we have shown that ML-SPERE and the IME method produce larger differences in emission rate estimates, though the results of Sect. A5 show that ML-SPERE generalizes beyond the spatial domain of training data."

R2.19 Fig. A3 a: Do plumes that extend to the edge of the domain incur a bias in either the IME or the ML-SPERE method?

C2.19 This is not the case for ML-SPERE or the IME method, and we have added a sentence at lines 284-285 to state this: "We find no significant change in bias for either method when applied only to plumes whose masks are truncated by the scene boundary."

R2.20 Line 298: Please note that Plewa et al. 2025 (https://doi.org/10.1016/j.rse.2025.115002), whom you cite above, discussed and found a solution for a low bias at high emission rates in their ML-based emission estimation method. The solution was to extend the range of emissions in the training dataset beyond emission rates that the trained model is later applied to. I think that solution could be applied to improve ML-SPERE as well - perhaps include that as an outlook for ML-SPERE.

C2.20 As discussed in C2.10, we believe that the choice of distribution for our training labels is the best fit for this problem. We cannot extend the training labels lower than we already have, as the lower bound of the training labels used in this study is already 0.5 $\mathrm{t\ /\ hr}$. As expressed in point C2.10, changing the distribution of the labels from our gamma distribution to a uniform distribution would potentially cause ML-SPERE to overestimate emissions from most plumes detected by TROPOMI using the pipeline of Schuit et al. (2023b) owing to regression dilution.

R2.21 Line 299: "greatly alleviated" - please be more specific (e.g. "reduced by 1/3", judging from the linear fit equations in Figures 3d vs A4d).

C2.21 We have clarified the sentence at lines 324-325, which now reads as follows: "This regression dilution for ML-SPERE is reduced by 1/3 when we restrict application to well-observed plumes (Fig. A4**d**), though no such improvement in estimate biases is found for the IME method (Fig. A4**c**)."

R2.22 Fig. 4a: Please double-check whether the OLS equation is correct. The intercept in the plot is at a slightly different position than in the equation (around -38 t/h instead of -44.21 t/h), and "44.21" looks like it could be a copy-paste error from the intercept shown in Fig.

4b (-4.21 t/h).

C2.22 We have double checked the equation and figure and believe them to be correct. The x-axis of Fig. 4 does not extend fully to 0 m / s, but rather to 0.25 m / s., so the intercept of the equation should line up with the intercept when the trend line is extended down to 0 m / s. The reason that the figure does not extent fully to 0 m / s is that there are no test set scenes with wind speeds below 0.25 m / s. It is an interesting coincidence that the intercepts in both OLS equations are so similar however.

R2.23 Line 303: "IME ... slightly overestimating them at high wind speeds". The figure that supposedly shows this (Fig. 4a) stops at 2 m/s - still a low wind speed. Also, the IME results in Fig. 4a don't show a bias at the "high" wind speed as claimed in the text. Please extend the figure to show the results for higher wind speeds. The same applies to Fig. 6b.

C2.23 We thank the reviewer for this comment and apologize for the lack of clarity. Our intent in the usage of "high wind speeds" was to refer to the upper portion of the wind speed distribution for detected TROPOMI methane plumes (i.e., above 2 m / s), rather than to high wind speeds in an absolute sense. In this regime, an overestimation is observed, although we agree this is not clearly conveyed in the original text or figure.

To address this, we have revised the text to focus on the more robust and clearly supported result: the underestimation at wind speeds below 2 m / s. The text at lines 325-328 now reads as follows: "For scenes with wind speeds below 2 m / s, residuals between ML-SPERE emission rate estimates and true plume emission rates show no correlation with scene wind speed. In contrast, residuals of IME estimates in this wind speed regime exhibit a strong positive correlation, with the IME method significantly underestimating emissions for these plumes." We have not extended the figures to higher wind speeds, as the majority of detections lie within the currently shown range and the key conclusions are captured there.

R2.24 Lines 308-310: Is this not a shortcoming of this particular effective wind speed parameterization, rather than of the IME method? It seems like it could be fixed.

C2.24 We agree, although we believe such improvements to the IME method implemented here and in Schuit et al. (2023b) is out of scope of this work. We have addressed a similar comment from Reviewer 1 in point C1.7.

R2.25 Lines 10, 339 and 455: In my opinion, the statement that "ML-SPERE agrees well with the inversion" sounds too vague and downplays discrepancies between the methods: In Fig. 5 and subsequent paragraphs, it is shown that Tropomi IME and Tropomi ML-SPERE are much closer to each other than either is to the Tropomi inversion - there is a large bias and low correlation between the results from either method. To me, the phrasing in abstract and conclusions sounds like there is a much bigger improvement in the agreement when using ML-SPERE instead of IME than there actually is. Please phrase these statements more carefully.

C2.25 We have phrased these statements you referenced more conservatively, which now read as follows:

1. (Lines 9-12) "Applied to TROPOMI observations of a 200-day well blowout in Kazakhstan, ML-SPERE typically shows agreement with TROPOMI inverse modeling results and TROPOMI IME estimates. Compared to TROPOMI IME estimates, ML-SPERE estimates show improved agreement with IME estimates derived from high-resolution point-source imagers."
2. (Lines 364-366) "In general, ML-SPERE estimates tend to be higher than the TROPOMI inverse modeling results presented in Guanter et al. (2024), although they typically agree within uncertainty and show slightly closer agreement than that between the TROPOMI inverse modeling and IME estimates."
3. (Lines 482-486) "When using ML-SPERE to quantify emissions for the 2023 Karaturun East blowout in Kazakhstan, we find that ML-SPERE estimates typically agree within uncertainties with IME and inverse analysis estimates from TROPOMI observations. Compared with TROPOMI IME estimates, ML-SPERE estimates show improved agreement with the high-resolution IME estimates. ML-SPERE estimates show a slightly reduced bias relative to TROPOMI IME estimates when compared against inverse modeling though scatter remains."

We also would like to point to text beginning on line 381 where we demonstrate that the TROPOMI inverse modeling estimates presented in Guanter et al. (2024) may be too conservative, i.e. biased low due to the double-penalty of mismatched plumes. The gap between ML-SPERE estimates and the inverse modeling results presented in Guanter et al. (2024) is significantly reduced when ML-SPERE is applied to the simulated plumes that were taken to be the "best match" to the observations in Guanter et al. (2024), and not to the observations themselves; this implies that there may be a mismatch between the simulations of Guanter et al. (2024) and the observations. Your comment R2.27 indicates that we did not make this fully clear, so we have also altered this section which is address in our reply C2.27.

R2.26 Line 345: "For days in the time series, ..." - This phrase sounds incomplete; did you mean to include a number of days here?

C2.26 Thank you for catching this error, we previously omitted a word and the sentence at lines 371-372 now reads as "For most days in the time series, the uncertainty in ML-SPERE predictions is dominated by uncertainty in the input wind data."

R2.27 Lines 358f: I don't fully understand this sentence. In line 358, it sounds like the metrics (R and bias) refer to ML-SPERE vs inversion results. In line 359, it says these numbers describe the agreement between ML-SPERE and a known truth. Which is it? Also, it says here that these results are "improved" compared to applying ML-SPERE to the actual data. But if I understand correctly, they refer to a subset of the cases from Fig. 5. So the reference point for the comparison (ML-SPERE applied to real data and inversion for this subset) is missing.

C2.27 We apologise for the confusion. As alluded to in comment C2.25, at this stage we switch from applying ML-SPERE to the TROPOMI observations of the Karaturun blowout, and move instead to applying ML-SPERE to the best-match simulated plumes from Guanter et al. (2024). This means that we can now take the simulated emission rates (i.e., the estimates) from Guanter et al. (2024) as ground-truth values and compare them to ML-SPERE estimates. In doing so, we find the gap is much more closed between the ML-SPERE estimates and simulated emission rates. IME estimates obtained from the same simulated plumes remain overestimated. We have altered the text at lines 381-385 to read as: "To investigate potential sources of bias between ML-SPERE and the inverse modeling estimates for these scenes, we apply ML-SPERE directly to the resampled simulated plumes from Guanter et al. (2024) that were identified as the best matches to the observations. Using these simulated plumes (without missing data or background methane) and their corresponding meteorology as inputs, ML-SPERE shows improved agreement with the simulated emission rates (i.e., the results shown in Guanter et al. (2024)), achieving a Pearson correlation of $R = 0.91$ and a substantially reduced mean bias of 15.4 $\mathrm{t\ /\ hr}$."

R2.28 Line 362: What do you mean by steady state emission and transport assumptions? Constant mean wind speed? Be specific. All other occurrences of "steady-state" in the text only refer to emissions.

C2.28 We have altered the sentence at lines 387-389 to be more specific and conservative, which now reads as "The remaining positive bias of 15.4 $\mathrm{t\ /\ hr}$ in the ML-SPERE estimates could be attributable to plume morphologies resulting from complicated wind fields or conditions which violate the steady-state emission assumptions represented in the training dataset."

R2.29 Line 362-364: I might have missed it, but why is it expected that IME is affected more strongly than ML-SPERE by violations of steady-state assumptions?

C2.29 The central assumption of the IME method is that wind speeds and an effective plume length are used to estimate a residence timescale, which is used to divide the estimated plume mass into an emission rate which really represents a time-averaged source strength prior to the satellite overpass. If emission variability allows methane to accumulate in the plume beyond steady-state levels, this can lead to an overestimation of the emission rate. Complicated wind fields producing curved, blob-like or otherwise complicated plume structures also inhibit the estimation of plume effective lengths, which similarly can misconstrue IME estimates. However, we have shown in Sect. A3 that ML-SPERE estimates are primarily informed by the alignment between gradients in methane abundance and wind vectors at the plume head, which means that ML-SPERE is less influenced by the morphology of plume tail than the IME method is. We have altered the text at lines 389-391 to reflect this (though please refer to the surrounding text in the manuscript for further context as we have altered this section in response to the reviewers' comments): "This interpretation is supported by the fact that the IME method (which assumes steady-state emission scenarios) produces estimates that are twice as large as those of the ML-based estimates when applied to the same simulated plumes from Guanter et al. (2024)."

R2.30 Lines 375f: "inverse modeling is generally considered the most accurate approach for estimating methane emissions from TROPOMI data when good plume matches can be obtained". See my comment to line 64f. I think the inversion might underestimate the true emissions for the reasons explained there.

C2.30 As with C2.5, we have altered the text at lines 400-404 to read as "Although the actual underlying emission rates are not known for this super-emitting blowout, the improved agreement between ML-SPERE and inverse modeling estimates (compared to IME method vs. inverse modeling agreement) is notable as inverse modeling is a far more information-rich method, incorporating three-dimensional wind fields and atmospheric transport processes to produce plume emission rate estimates (albeit at a substantially higher computational cost than other methods)."

R2.31 Lines 384-386: "ML-SPERE estimates tend to exceed those of the IME method at low IME-estimated emission values, but are generally lower at higher IME-estimated emission values". Add a figure in the appendix that demonstrates this statement? Fig. 6a kind of has this I guess, but I find it hard to see. A plot of the difference between ML-SPERE and IME vs either estimate could show this better.

C2.31 We have changed Fig. 6**a** to a log-log scale to make this clearer, and have altered our text at lines 411-415 to read as "ML-SPERE estimates tend to exceed those of the IME method at low IME-estimated emission rates (i.e., below 10 $\mathrm{t\ /\ hr}$). This behavior may reflect the regression dilution discussed in Sect. 3.1 and shown in Fig. 3**c**. Additionally, it may indicate underestimation by the IME method for these plumes due to the wind-speed bias described in Sect. 3.1, as discussed further below. Emission estimates for TROPOMI plumes below 10 $\mathrm{t\ /\ hr}$ should be interpreted with caution, as this lies near the conventional detection threshold."

R2.32 Lines 401-407: The first reason for potential overestimation given in A8 - regression dilution - is not mentioned here. Why not include it?

C2.32 At the lines you reference, we discuss the discrepancy between ML-SPERE and IME estimates for pipeline emissions in northern Russia. We show in Fig. A9 that in this region, ML-SPERE estimates are on average higher than those produced via the IME method. However, the estimates produced via the IME method are already much larger than the mean emission rate of our training dataset labels. If we take the IME estimates as a piece of evidence that the true emission rates of these pipeline emissions are larger than the mean of our training dataset (which we admit is an assumption), then regression dilution would cause ML-SPERE to underestimate these emissions, not overestimate them. It is perhaps more likely for these plumes that extreme gradients in methane abundance are causing ML-SPERE to produce very large estimates.

R2.33 Line 406: "especially if the plume has detached" - Why does detachment of the plume cause IME errors? Is it simply a strong deviation from the steady-state assumption or is something else going on?

C2.33 We thank the reviewer for this question. When a plume is detached from its source, the steady-state assumptions underlying the IME method are strongly violated. In such cases, the delineated plume mass no longer reflects a balance between a constant emission rate and wind-driven transport; instead, it represents an accumulation of emissions released at earlier times under a time-varying emission profile (i.e., a transition from no emission to active emission and back to no emission). In northern Russia, these bursts of pipeline emission often appear as small blobs of high methane enhancements that may only be a few TROPOMI pixels in size (see De Jong et al. (2025) Fig. 3 for a similar example in Kazakhstan). We have altered the sentence in the text at lines 433-435 to make this clearer: "Such transient, non steady-state emissions may also not be well estimated by the IME method (especially if the plume has detached from the source and is being transported downwind, as this violates the steady-state assumptions underlying the method)."

R2.34 Line 407: "the total plume mass may provide a more important characterization" more important than what and for what?

C2.34 We have altered this sentence at lines 435-438 to reflect the fact that emission is no longer ongoing and that a total methane mass emitted by the event can be calculated, which may be more relevant than attempting to calculating an instantaneous emission rate. "For these plumes, the total plume mass may provide a more relevant characterization than estimated instantaneous emission rate as emissions are no longer ongoing (De Jong et al., 2025). For ongoing or persistent events, the observed mass only provides a lower boundary for the total emitted mass, as mass is still being emitted at the time of observation."

R2.35 Lines 13f, 409 and 634-637: Could you please confirm / elaborate on the conclusion that IME estimates could be too high for diffuse sources? I tried following this argument with pen and paper and arrived at the opposite conclusion.

*Consider a simple "diffuse" source consisting of two identical point sources, Qsingle, in adjacent pixels of the swath, and that their two plumes are right next to each other but don't overlap - forming a single, broader plume from a "diffuse" source. The true emission rate is 2*Qsingle. The combined IME is twice the single IMEs, and the combined length is sqrt(2) times the length of the single plumes (because the combined area is twice the single areas and L=sqrt(A)). If I compute Q from the combined plume, I get $u_eff*(2*IMEsingle)/(sqrt(2)*Lsingle) = sqrt(2)*Qsingle$, an underestimation of the true emission rate of 2*Qsingle. Intuitively, the reason for the underestimation is that the broadened plume has a "longer" length L (because it's computed from the area) than a point source with the same emission rate would have. So IME misinterprets the more diffuse source as a longer residence time.*

From these considerations, I conclude that IME has a risk to underestimate emissions from diffuse sources - the opposite of what the text says and what we seem to see in the wetland results in Fig. 6c, where the IME estimate is higher than ML-SPERE. Perhaps I made an error in my simplified case. But perhaps ML-SPERE also underestimates area sources, but more severely than IME?

C2.35 To simplify our analysis we have altered the text at the following locations to make clear that the main point is that the IME method (as well as ML-SPERE) was not constructed for use with area sources, and the application of either method under these circumstances must be treated with caution.

- (Lines 12-16) "Global spatial patterns of methane emissions inferred from ML-SPERE and the IME method for all super-emitters found by TROPOMI in 2021 are broadly consistent, with notable regional differences in northern Russia (where transient pipeline emissions may not be well characterized by either method), the Congo Basin (where area source emissions may not be well characterized by either method), and southeastern Australia (where IME estimates are potentially negatively biased owing to predominantly low wind speeds)."
- (Lines 438-440) "In contrast to pipeline emissions in northern Russia, methane enhancements in the Congo Basin have very large spatial extents and may be more representative of area-source emissions than (TROPOMI-scale) point-source emissions on which ML-SPERE was trained and the IME method was calibrated."
- (Lines 663-665) "Consequently, ML-SPERE may be less well suited for application to these scenes, though it is important to recognize that the IME method was not designed to be applied to area-source emissions. Both methods may struggle to characterize emissions in this region."

References

Bruno Jack H., Jervis Dylan, Varon Daniel J., Jacob Daniel J. U-Plume: automated algorithm for plume detection and source quantification by satellite point-source imagers // Atmospheric Measurement Techniques. V 2024. 17, 9. 2625–2636.

Copernicus Atmospheric Monitoring Service . CAMS Methane Hotspot Explorer. II 2025.

De Jong Tobias A., Maasakkers Joannes D., Irakulis-Loitxate Itziar, Randles Cynthia A., Tol Paul, Aben Ilse. Daily Global Methane Super-Emitter Detection and Source Identification With Sub-Daily Tracking // Geophysical Research Letters. IV 2025. 52, 8. e2024GL111824.

Guanter Luis, Roger Javier, Sharma Shubham, Valverde Adriana, Irakulis-Loitxate Itziar, Gorroño Javier, Zhang Xin, Schuit Berend J., Maasakkers Joannes D., Aben Ilse, Groshenry Alexis, Benoit Antoine, Peyle Quentin, Zavala-Araiza Daniel. Multisatellite Data Depicts a Record-Breaking Methane Leak from a Well Blowout // Environmental Science & Technology Letters. 2024. 11, 8. 825–830. _eprint: https://doi.org/10.1021/acs.estlett.4c00399.

Irakulis-Loitxate Itziar, Guanter Luis, Maasakkers Joannes D., Zavala-Araiza Daniel, Aben Ilse. Satellites Detect Abatable Super-Emissions in One of the World's Largest Methane Hotspot Regions // Environmental Science & Technology. II 2022. 56, 4. 2143–2152.

Jongaramrungruang Siraput, Thorpe Andrew K., Matheou Georgios, Frankenberg Christian. MethaNet – An AI-driven approach to quantifying methane point-source emission from high-resolution 2-D plume imagery // Remote Sensing of Environment. II 2022. 269. 112809.

Joyce Peter, Ruiz Villena Cristina, Huang Yahui, Webb Alex, Gloor Manuel, Wagner Fabien H., Chipperfield Martyn P., Barrio Guilló Rocío, Wilson Chris, Boesch Hartmut. Using a deep neural network to detect methane point sources and quantify emissions from PRISMA hyperspectral satellite images // Atmospheric Measurement Techniques. V 2023. 16, 10. 2627–2640.

Plewa Thomas, Butz André, Frankenberg Christian, Thorpe Andrew K., Marshall Julia. Improvements of AI-driven emission estimation for point sources applied to high resolution 2-D methane-plume imagery // Remote Sensing of Environment. XII 2025. 331. 115002.

Radman Ali, Mahdianpari Masoud, Varon Daniel J., Mohammadimanesh Fariba. S2MetNet: A novel dataset and deep learning benchmark for methane point source quantification using Sentinel-2 satellite imagery // Remote Sensing of Environment. IX 2023. 295. 113708.

Schuit Berend J., Maasakkers Joannes D., Bijl Pieter, Mahapatra Gourav, Berg Anne-Wil Van den, Pandey Sudhanshu, Lorente Alba, Borsdorff Tobias, Houweling Sander, Varon Daniel J., McKeever Jason, Jervis Dylan, Girard Marianne, Irakulis-Loitxate Itziar, Gorroño Javier, Guanter Luis, Cusworth Daniel H., Aben Ilse. Dataset: all TROPOMI detected plumes for 2021. [Schuit et al. 2023: Automated detection and monitoring of methane super-emitters using satellite data]. VI 2023a.

Schuit Berend J., Maasakkers Joannes D., Bijl Pieter, Mahapatra Gourav, Berg Anne-Wil van den, Pandey Sudhanshu, Lorente Alba, Borsdorff Tobias, Houweling Sander, Varon Daniel J., McKeever Jason, Jervis Dylan, Girard Marianne, Irakulis-Loitxate Itziar, Gorroño Javier, Guanter Luis, Cusworth Daniel H., Aben Ilse. Automated detection and monitoring of methane super-emitters using satellite data // Atmospheric Chemistry and Physics. IX 2023b. 23, 16. 9071–9098.

United Nations Environment Programme . An Eye on Methane 2025: From Measurement to Momentum. X 2025.

Varon Daniel J., Jacob Daniel J., McKeever Jason, Jervis Dylan, Durak Berke O. A., Xia Yan, Huang Yi. Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes // Atmospheric Measurement Techniques. X 2018. 11, 10. 5673–5686.