We thank the reviewer for their constructive comments and insightful suggestions, based on which we have thoroughly revised the manuscript. The main revisions are summarized below, with detailed, point-by-point responses provided in the following sections. Our responses are shown in black, while the original comments are reproduced in blue italics.

1. **Expanded false positive and false negative analysis.** We added a clear definition of a plume in the Methods and conducted a more comprehensive error analysis in the Results. We identified background interference and diffuse enhancements as the primary sources of false positives, and small plumes and low–flux emissions coexisting with strong sources as the main causes of false negatives, with additional challenges coming from merged plumes. These findings motivate an expanded discussion of future work in Discussion, including targeted wavelet denoising in high-probability subregions, adaptive and region-specific parameter tuning, and the development of classifiers for false detections.
2. **Broader comparison with alternative methods.** We extended the evaluation to include additional denoising approaches (Gaussian filtering and median filtering) and alternative wavelet families (Daubechies, Symlets, and biorthogonal wavelets), both assessed based on detection performance. We also added a comparison with a machine learning approach in terms of computational efficiency.
3. **Quantitative assessment using a probability of detection framework**. We incorporated a probability of detection (P_d) analysis following Manninen et al. (2026), enabling a more rigorous comparison of detection limits across detection methods and observing systems. Results indicate that the wavelet method generally outperforms DI thresholding by detecting more plumes, including those with lower emission rates. For MethaneAIR, the wavelet method achieves $P_d \approx$ 0.93 and 0.42 at wind speeds of 1 m/s and 10 m/s, respectively, for emissions of 1000 kg/h; for MethaneSAT, $P_d \approx$ 0.93 and 0.43 under the same wind conditions for emissions of 2000 kg/h.

Major comments:

The study uses CH4 maps derived from hyperspectral data. As hyperspectral data contains much more redundancy of information in the spectral bands, which is beneficial for denoising, the authors should either motivate why they apply the denoising only to the CH4 maps or use a more sophisticated denoising method such as 3D wavelets or HyRes and compare it to the current results using 2D wavelets.

Response:

We thank the reviewer for highlighting the potential advantages of hyperspectral denoising methods such as 3D wavelets and HyRes.

In this study, our denoising is applied to the retrieved CH_4 concentration maps rather than to the original hyperspectral radiance data, motivated by both our methodological scope and practical considerations. The CH_4 maps used here are already the result of a complex retrieval pipeline that incorporates spectral information, and thus represent a physically meaningful, reduced-dimensional data product. Therefore, applying denoising at this stage is much more computationally efficient than operating on the full spectra, and it allows the method to be applied directly to Level-2/3 products without modifying upstream processing chains. In contrast, processing full radiance data would introduce additional computational and implementation complexity.

We agree that incorporating hyperspectral-domain denoising is an interesting research idea to explore. However, such approaches would fundamentally shift the scope of this work toward retrieval-level

optimization. Our goal here is instead to demonstrate that even a simple, computationally efficient 2D wavelet-based denoising applied to CH_4 maps can significantly enhance detection performance.

To address the reviewer's suggestion, we have added the studies mentioned by the reviewer to our literature review, and we have added an explanation in the Methods to clarify this design choice.

Introduction: "In remote sensing applications, wavelet transform has been applied primarily for object detection and removal, such as identifying and eliminating haze and clouds from optical satellite imagery (Schneising et al., 2023). *Wavelet-based methods have also been extended to hyperspectral data (Rasti et al., 2018), including full-spectra denoising approaches (e.g., 3D wavelets), which leverage spectral redundancy for noise reduction (Rasti et al., 2017)*."

Section 2.2.2: "*We focus on denoising retrieved CH_4 concentration maps rather than the full hyperspectral radiance data, as operating on full spectra would require modifying the retrieval pipeline and is beyond the scope of this study.* We adapt an approach used in medical image processing, providing its first implementation in atmospheric studies (Hüpfel et al., 2021)."

The goal of remote sensing of trace gases is not only to detect sources but also to quantify their emissions. Therefore, the authors should also quantify the impact of the proposed denoising approach (including hyperparameter choices) on the resulting emission estimates and detection limits. Since this requires access to ground truth, a suitable approach would be to use synthetic plume images and add noise with different characteristics (e.g., Gaussian and non-Gaussian). This would allow a systematic assessment of whether and to what extent the denoising step biases or alters the inferred emission rates.

Response:

We thank the reviewer for this important comment. We agree that quantifying the impact of the proposed approach on emission estimates and detection limits is a critical aspect of assessing its performance. In fact, a closely related analysis has been conducted by our team in a separate study (Manninen et al., 2026, https://doi.org/10.5194/egusphere-2026-115), where plume detection performance is systematically evaluated using both simulated and controlled-release datasets.

This study develops a unified probability of detection (P_d) framework to compare plume detection performance across different observing systems (including MethaneAIR and MethaneSAT) and detection algorithms (including the wavelet method and the DI thresholding method). The framework introduces a nondimensional "observability" predictor that combines emission rate, wind speed, pixel size, and gas concentration noise, and maps it to detection probability using logistic regression (here observability $=\log\left(\frac{q}{\sqrt{a}nu}\right)$, q = emission rate per unit area, a = pixel area, n = noise amplitude, u = windspeed). The analysis is based on ~80,000 synthetic plumes generated using Weather Research and Forecasting Large Eddy Simulations (WRF-LES), spanning a wide range of atmospheric conditions and noise characteristics, as well as ~62,000 scenes derived from controlled-release experiments using image processing methods.

The results show that the wavelet method generally outperforms the DI thresholding method by detecting more plumes overall, particularly at lower emission rates, and by being more robust to real-world observations with missing data. Quantitatively, the wavelet method achieves higher P_d than DI above an observability of ~1.3, with the largest performance gains at intermediate observability (~2.5, a SI figure with P_d as a function of observability is provided). At very low observability, the wavelet method shows

slightly lower P_d, likely due to suppression of weak signals during multi-scale denoising. This behavior is consistent with the patterns observed in our MethaneAIR and MethaneSAT case studies.

The companion study also provides quantitative detection limits under different conditions. For example, for MethaneAIR, the wavelet method achieves P_d of 0.93 and 0.42 at wind speeds of 1 m/s and 10 m/s, respectively, for emissions of 1000 kg/h. For MethaneSAT, comparable P_d values (0.93 and 0.43) are achieved for emissions of 2000 kg/h under the same wind conditions. A summary of detection probabilities across different emission rates and wind speeds is provided in Table 3.

We acknowledge that this analysis evaluates the performance of the full detection framework and does not explicitly isolate the effect of the denoising step. In our approach, however, the wavelet-based denoising serves as a foundational step for plume mask generation, which directly determines detection outcomes. Subsequent filtering steps are primarily designed to remove false detections and do not fundamentally alter the underlying plume structures identified during denoising (the number of detections at each processing step is now shown in Table 2). As such, the overall detection sensitivity presented here largely reflects the effectiveness of the denoising-based mask generation (discussion on this is added in Section 3.1).

We have added a subsection, Section 4.2 “Probability of detection” under the Discussion to clarify this point. We also highlighted the value of future work using simulation data and controlled release observations to further quantify the role of denoising in emission estimation and detection limits.

Specific comments:

L11ff: How much lower is the effective detection limit with your method?

Response: We kindly refer the reviewer to our response to the previous comment.

L12: There are many more disadvantages of ML methods that could be listed here like the lack of uncertainty propagation, regression to the mean, often unrealistic training data etc.

Response: We thank the reviewer for providing these additional perspectives. Due to word limits, we have kept the abstract unchanged; however, we have incorporated these points into the Introduction.

“However, these ML models still have several limitations..... *Additional limitations include limited interpretability, uncertainty propagation, and biases toward the mean*.”

L13 & L70: What are "MethaneSAT purposes"?

Response: By “MethaneSAT purposes,” we refer to the plume detection objectives of the MethaneSAT mission. We have revised the wording to clarify this point.

“*It is designed to be readily adaptable to multiple aircraft and satellite platforms and to be applicable to other trace gases that exhibit plume-like structures associated with discrete sources*.”

L35ff: One important paper to cite would also be Kuhlmann et al., 2024 (https://gmd.copernicus.org/articles/17/4773/2024/)

Response: We thank the reviewer for the suggestion, and we have added this citation in the revised manuscript.

"The most common approach to plume masking involves simple thresholding, typically based on concentration values (Duren et al., 2019), percentiles (Sánchez-García et al., 2022), standard deviations (Chulakadabba et al., 2023), significance tests (Varon et al., 2018), *or signal-to-noise ratio (Kuhlmann et al., 2024)*."

L73 & 85: please indicate the along and across-track resolution

Response: We have explicitly indicated the along and across-track resolution in the manuscript.

"With (MethaneSAT's) ..... spatial sampling *(110 m across-track × 400 m along-track)*"

"(MethaneAIR) providing ..... fine spatial resolution *(5 m across-track × 25 m along-track)*"

L97: As far as I can follow, the pre-processing step was not part of the processing in Hüpfel et al., (2021). Please provide more details about why this is needed in the current study but was not in Hüpfel et al.

Response: We thank the reviewer for pointing this out. In medical imaging applications such as those in Hüpfel et al. (2021), the targets are typically more spatially coherent and exhibit higher contrast relative to noise, which makes the separation of signal and noise more straightforward. In contrast, methane plume signals generally have a much lower signal-to-noise ratio, a heterogeneous background, and additional retrieval artifacts that introduce structured noise. These factors make direct wavelet-based separation more challenging in our application. Therefore, we introduce an additional pre-processing step to reshape strong methane enhancements into more spatially coherent, low-frequency structures, which facilitates more effective separation from high-frequency noise.

We have revised the manuscript to explicitly state the difference between two applications that motivates the pre-processing.

"We adapt an approach used in medical image processing not previously implemented in atmospheric studies (Hüpfel et al., 2021). *In Hüpfel et al., (2021), targets are more spatially coherent and exhibit higher contrast relative to noise than methane concentration maps. Therefore,* our method begins with *an additional* pre-processing step *where all pixels above a defined threshold are assigned a uniform value*, converting strong signals to low frequencies to help separate them from high-frequency noise."

L99: What you are doing here is basically the 99% confidence interval. Please make this clearer to the reader.

Response: We thank the reviewer for this comment, and we have clarified that the thresholds correspond to an one-sided Gaussian cutoff at a specified percentile.

"For MethaneAIR, the threshold is uniform across the scene, determined as the mean value of the whole image scene plus a scaling factor (value: 2) times the standard deviation of the whole scene, *corresponding to approximately the 97.725th percentile under a Gaussian assumption*; and for MethaneSAT, the threshold is adjusted in varying local background, defined as the mean value for a

subregion (4.5×4.5 km^2) of the image plus a scaling factor (value: 1.75) times the standard deviation of this local area, *corresponding approximately to the 95.5th percentile under a Gaussian assumption*."

L104: The authors state that the hyperparameters "were chosen to maximize plume detections while minimizing false positives". This risks overfitting to the evaluation dataset (e.g., similar sensor properties, surface characteristics etc.). Therefore, this issue should either be addressed or discussed.

Response: We thank the reviewer for raising this point. We acknowledge that selecting parameters to maximize plume detections while minimizing false positives can introduce a risk of overfitting to the evaluation dataset. The optimal parameter values may also depend on factors such as instrument noise characteristics, surface heterogeneity, and atmospheric conditions. Therefore, as more data become available, further evaluation across a wider range of regions and observing conditions will be important to assess generalizability and refine parameters.

In addition, the framework allows for flexibility in parameter selection, and future work could explore adaptive or region-specific tuning strategies to improve performance under varying conditions. We have added discussion in the Discussion section to clarify this point.

Section 4.1.2: "These parameters were chosen to maximize plume detections while minimizing FPs, and are readily tuned for different observing platforms. *Selections should be made to minimize overfitting to the datasets being considered, as discussed further in Section 4.1.2.*"

L122: How did you arrive at this value for L? Did you perform sensitivity analysis? How do the estimated emissions change when changing L? It would be good to include this in your sensitivity analysis.

Response: We thank the reviewer for raising this important point. We agree that the choice of the decomposition level L should be justified in more details.

In response, we conducted a sensitivity analysis of L using three MethaneAIR scenes and three MethaneSAT scenes (the same scenes used for other parameter sensitivity analysis). We evaluated six candidate values corresponding to fractions of the maximum possible decomposition level: $L = L_{max} \times [\frac{1}{8}, \frac{3}{8}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, 1]$. For each value, we quantified the number of true and false plume detections.

The results show a consistent trend in which both true and false detections generally decrease as L increases. Smaller values of L (e.g., $1/8 L_{max}$) may retain more high-frequency components, resulting in a large number of false positives. In contrast, larger values (e.g., L_{max}) tend to over-smooth the data and suppress plume signals, leading to missed detections. Across all tested scenes, $L = 1/2\ L_{max}$ provides the best balance between detection sensitivity and false positive control. Based on these results, we selected $L = 1/2\ L_{max}$ as a robust and empirically justified choice. We have added this sensitivity analysis to the SI Table S2 and Table S12 and clarified the rationale for this selection in the main text.

"In this process, the decomposition level L is a key parameter. *Smaller values of L may retain more high-frequency components, resulting in a large number of false positives. In contrast, larger values tend to over-smooth the data and suppress plume signals, leading to missed detections (SI Table S2 and Table S12).* To balance these effects, we set L at half of its maximum possible value."

L124: Please explain why this additional denoising needed if your previous wavelet transform is supposed to filter out the high frequency noise (L118)? Especially since this step was not performed in Hüpfel et al., (2021).

Response: We thank the reviewer for the comment. Unlike in Hüpfel et al. (2021), plume signals and noise in our application are not cleanly separable in the frequency domain as both can occupy overlapping scales due to background heterogeneity and retrieval artifacts. As a result, some residual noise may remain in the reconstructed image after the initial wavelet filtering. The additional soft thresholding step introduces a non-linear filtering that further suppresses the residual noise while preserving the main plume structures. This step is particularly useful for attenuating weak noise that survives the earlier processing stages.

We also acknowledge that soft thresholding may introduce some degree of over-smoothing, particularly for very weak plumes. However, our parameter choices aim to balance noise suppression and signal preservation, and the improved detection performance suggests that this trade-off is beneficial overall. We have added clarification in the Methods section to explain the role of this step and its associated limitations.

"Finally, we apply soft thresholding wavelet denoising to the processed image for further denoising. *This process begins with a wavelet decomposition, followed by shrinking the wavelet coefficients based on their magnitude, such that the coefficients lower than a defined threshold (usually zero) are reduced to the threshold value, while larger coefficients are preserved (Donoho and Johnstone, 1998). As a non-linear filter, it helps suppress small noise-associated coefficients that persists from earlier processing stages, while retaining larger coefficients most likely to represent meaningful features. However, this approach may introduce some degree of over-smoothing, particularly for very weak plumes.*"

L149: Please provide values of high and low ratios. Also provide the thresholds used.

Response: We have added high and low hotspot ratio thresholds into the manuscript.

"We define a hotspot filtering criterion named the hotspot ratio, which is the ratio of the hotspots pixel count to the total pixel count of the mask. Masks with a low hotspot ratio *(threshold: 0.002)* are discarded, while those with a high hotspot ratio *(threshold: 0.03)* are accepted without further filtering. Masks with a hotspot ratio in between are further evaluated by the latter filters."

L165ff: At the scale of MethaneSAT, using a coarsely resolved wind data product such as HRRR or GFS might work well, especially if there is little topography as it was the case for the data presented. However, at smaller spatial scales (such as for MethaneAIR) and in more complex terrain, filtering your data based on wind direction might become problematic as the influence of turbulence grows larger (see e.g., https://amt.copernicus.org/articles/19/333/2026/). Please discuss this limitation in the discussion section.

Response: We thank the reviewer for highlighting this important limitation. We agree that the use of coarsely resolved wind products may become problematic at finer spatial scales or in complex terrain due to increased flow variability and turbulence. In our current implementation, we partially mitigate this issue by adopting a relatively wide wind-direction buffer (±55°), although mismatches can still occur in some cases. We have added discussion in the Methods to acknowledge this limitation and to outline potential improvements.

Section 2.2.4: "*We note that the wind-direction filtering relies on external meteorological data, which may be too coarsely resolved for applications at finer spatial scales or in complex terrain with higher local flow variability and turbulence influence. This may introduce uncertainty in plume alignment and potentially lead to incorrect filtering. Our current wind-direction buffer is relatively wide; however,*

mismatches may still occur in some cases. Future work could improve upon this by incorporating plume morphology to infer or refine wind direction independently of external meteorological data."

L188: Please explain which wind speed product you used for the emission quantification (e.g., 10m wind speed, PBL average, effective wind speed and how you calculated it.)

Response: We have added the specific wind product we used (80m wind speed for both HRRR and GFS) into the manuscript.

"Wind speed *and direction* from the meteorological product *at 80m height (HRRR or GFS, selected to represent the surface layer windspeed)* was used for this calculation, and for sufficiently elongated plumes (eccentricity > 0.87), the wind direction was rotated to match the angle of the major axis of the observed plume."

L235ff: Please be more concrete about the performance of your algorithm: How many low concentration plumes were filtered out and how many remained?

Response: We thank the reviewer for this suggestion. We now have added proportions of each false positive and false negative category into the Results section. For example, diffuse enhancements ("low concentration plumes") discussed here account for 42% of the total false positives produced by the wavelet method. In addition, we have added new tables of three MethaneAIR scenes and three MethaneSAT scenes in Table 2 that report the number of true and false detections after each processing step. These tables provide examples to further illustrate how many false detections are filtered and how many remained.

L263: You could also mention that the longer the plumes, the more the assumption of steady wind conditions are violated

Response: We thank the reviewer for bringing this point, and we have added it into the manuscript.

".....restricting the masks to the plume segment within 5–10 km of the source generally produces more accurate flux estimates, because expanding the masks farther downwind increases the chance of including nearby sources in the DI growing boxes, thereby biasing flux calculations. *Expanding the masks can also increase violations of the steady-wind assumption due to spatial and temporal wind variability*."

L274ff: Please elaborate why the application of the wavelet transform to subregions would improve the sensitivity of your method.

Response: We thank the reviewer for this question. The application of the wavelet transform to subregions improves sensitivity mainly because we can lower detection thresholds in regions with previously detected plumes or known oil and gas infrastructure. This allows us to capture weaker plume signals that may be missed in full-scene processing without substantially increasing false positives. In addition, applying the wavelet method to subregions helps reduce the impact of spatial heterogeneity in background, thus reduce the number of false positives. We have added explanations to the Discussion to make it clearer to readers.

Section 4.3: "We have applied the wavelet to the whole scene with no prior information of point source locations. Looking ahead, we plan to further improve its sensitivity to plumes by applying targeted wavelet denoising to subregions where plumes have been identified in previous top-down surveys observations, or where known oil and gas infrastructure exists. *This could include a growing dataset of*

past observed plume locations by MethaneAIR and MethaneSAT, as well as integration with external datasets such as the methane point source database from the International Methane Emissions Observatory (IMEO)(UNEP-IMEO, 2023), and the Oil and Gas Infrastructure Mapping database (OGIM) (Omara et al., 2023). It would allow the use of lower detection thresholds in high-probability emission areas, so that we can capture weaker plumes that may be missed in full-scene processing. It also reduces false positives due to lower background heterogeneity within subregions."

L285: Please be more concrete about the performance of your method: Add numbers to the text or insert a confusion matrix or a summarising table that show how well your method performed compared to the previous results (e.g., TP, FP, FN, detection limits etc.)

Response: We thank the reviewer for the suggestion. We have added two new tables to provide a more quantitative evaluation of method performance. The first (Table 1) summarizes detection results for the wavelet and DI methods across all MethaneAIR and MethaneSAT datasets, reporting true and false detections to enable direct comparison. The second (Table 2) shows the cumulative impact of each processing step on true and false detections for representative scenes.

Technical corrections

L6 & 92 & 214 & 253 etc.: "Plume signals" sounds a bit off. Use plumes or CH4 enhancements instead.

Response: Changed "Plume signals" to "plumes".

L14 & 45: "concentration imagery" sounds a bit off. Use concentration maps or CH4 enhancement maps instead.

Response: Changed "concentration imagery" to "concentration maps".

L33: Not "plume detection methods" but "plume detection"

Response: Changed "plume detection methods" to "plume detection".

In my opinion, Figure 1 is quite unintuitive. Please set the label as titles, add colourbars below each subplot, make the figure caption more descriptive and separate the upper and lower row as the lower row is only supposed to show the processing used to obtain figure (c).

Response: We thank the reviewer for this helpful suggestion. We have converted subplot labels to titles, put colorbars horizontally below each subplot, and separated upper and lower rows, each added with an explicit row title. We have also expanded the figure caption to fully describe each panel.

Figure 1 caption: "Wavelet denoising workflow. *The figure is organized into two rows to distinguish the data processing pipeline from the underlying wavelet transform operations. Top row (left to right): (a) original input image, (b) pre-processed input image, (c) reconstruction from wavelet transform with approximation coefficients removed, (d) residual image obtained by subtracting (c) from (b), and (e) final denoised image after applying soft-thresholding wavelet denoising to (d). Bottom row: illustration of the 2D discrete wavelet transform workflow, showing (i) decomposition into wavelet coefficients, (ii)*

modification of coefficients by removing the approximation components, and (iii) inverse transform reconstruction used to generate (c)."

Please fix the tick labels in Figure 2, remove the note from the figure and put it into the figure caption, remove the text in the upper right and set it as title

Response: Following the reviewer's suggestion, We have standardized the spacing of the x- and y-axis ticks and moved the notes from the figure to the captions.

Figure 2 caption: "MethaneAIR plumes stacked histogram. Most of the plumes found only by the wavelet method (yellow) lie in lower flux ranges. Most frequently observed flux rates moved from 600-1000 kg/h to 200-600 kg/h when added with additional plumes found by the wavelet method. *The total plume count from two methods is 600.*"

L221: Please use a different expression, e.g. homogeneity

Response: Changed "cleanliness" to "homonegeity".

The plume outlines in Figure 6 are hardly visible for people with colour vision deficiency (e.g., with Achromatopsia, see https://www.color-blindness.com/coblis-color-blindness-simulator/). Please use more suitable colour.

Response: Changed plume outline colors from red to white to improve contrast and accessibility.