EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-771-RC2, 2022

Comment on egusphere-2022-771

Anonymous Referee #2

Referee comment on "Detecting and quantifying methane emissions from oil and gas production: algorithm development with ground-truth calibration based on Sentinel-2 satellite imagery" by Zhan Zhang et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-771-RC2, 2022

The manuscript by Zhang et al. deals with methane plume retrievals with the Sentinel-2 satellite mission. They use methane concentration enhancement maps derived from Sentinel-2 data over controlled methane releases to constrain free parameters in the retrieval and emission rate estimation algorithms. They show that the Sentinel-2 detection and quantification of methane plumes from those controlled releases improves after model calibration with the same in situ data.

In my opinion, the research discussed in this manuscript must be of interest to the methane remote sensing community, especially considering the recent and rapid development of satellite-based high-resolution methane mapping methods. Also, the topic fits perfectly in AMTD, where the first paper on the use of Sentinel-2 for methane mapping (Varon et al., 2021) was published.

On the other hand, I have some major concerns with the overall purpose and some technical details of this work. In particular, I am not sure about the value of calibrating the algorithms with ground truth in this case. Is there any hope that they can be extrapolated to other sites or even seasons at the same site? I would say no. The algorithm parameters that they are optimizing are strongly acquisition dependent. For example, the thresholds accounting for outliers and false positives are driven by surface characteristics (homogeneity, stability). The finding that 12 dates are optimal for the multitemporal method wouldn’t apply to a site with changing vegetation covers, for which a configuration with one recent reference acquisition would be better than with a combination of 12 of them. Also, the thresholds used to filter out outliers should depend on the heterogeneity of the site. The retrieval noise, and hence the spatial extent of the plume, will also depend on the surface heterogeneity.

This extrapolation question would also apply to the proposed two-step method to improve emission quantification. Is this relevant to a wide community if data from a controlled

methane release are needed as input?

I think that the authors should show that the estimated model parameters can be applied outside this particular experiment for this work to be relevant. I don't think that the no false-negative test in Fig. 8 is a proper assessment of the model extrapolation that I am asking for (no emission to evaluate, and acquisition conditions for site B are very close to those of site A).

Perhaps the authors could run tests of how those thresholds perform for other sites, especially those with a more complex surface such as the US sites included in Ehret et al. https://pubs.acs.org/doi/10.1021/acs.est.1c08575?

In addition, I think it should it be possible to use this nice dataset to investigate possible approaches for automatic estimation of the outlier filtering threshold and the plume definition threshold. Those thresholds should be based on scene-based noise/heterogeneity estimates, such as n-sigmas above the retrieval noise level. Perhaps the authors could come up with approaches to estimate those parameters from the methane enhancement maps and cross-compare the results with the values derived from the model calibration presented in the manuscript. Examples of such threshold estimation approaches can be found in Ehret et al. (Background Estimation).

Other comments:

Sec 3.1, List of steps to improve plume detection: most of those steps (clear-view overpasses, normalization, removal of outliers, multiple reference data) are relatively obvious and already included in existing algorithms (e.g. Ehret et al., Gorrono et al.). Because of this, I am not sure that the methodology presented in this manuscript deserves to be presented as a new method, including an own acronym.

L2. (and L159) "performance validation by calibration" – not sure what this means

L288 – Not sure about statements on thresholds and detection accuracy: comments might apply to the avoidance of false negatives, but not the occurrence of false positives (which is actually the main difficulty for plume detection in real detection scenarios).

Units: a space is missing between numbers and units, such as in "30m"

L145. Please, check citations.

Shouldn’t this preprint on the Stanford methane release experiment be cited https://eartharxiv.org/repository/view/3465/?