EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-924-RC1, 2022

Comment on egusphere-2022-924

Anonymous Referee #1

Referee comment on "Using a deep neural network to detect methane point sources and quantify emissions from PRISMA hyperspectral satellite images" by Peter Joyce et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-924-RC1, 2022

Joyce et al. present a computer vision framework for detecting, masking, and quantifying methane point source plumes in PRISMA satellite data. Their method is successful in application to both synthetic (training/validation) and real satellite observations and shows great promise for application to increasingly large satellite methane datasets to rapidly identify and quantify methane point sources. The paper is creative, well-written, and a valuable contribution to the growing body of literature on space-based monitoring of methane emissions. It is well-suited for publication in AMT. I recommend accepting the paper for publication after the authors address the comments below and those of the other referee(s). My main suggestion is to better balance the discussion of strengths and weaknesses of the machine learning approach compared to classical physics-based methods for plume detection/quantification.

Comments

- L. 21: 1 km^2 seems quite large for point sources except landfills. Point sources are usually on the order of m^2. Are you referring to the area of a detected plume, or just trying to include a range of source types?
- L. 23: Classical methods are certainly time-consuming, but are they inaccurate? I don't think so, e.g., see Sherwin et al. (2022): https://eartharxiv.org/repository/view/3465/
- L. 46: One of Daniel Cusworth's papers would be an appropriate reference for the temporal emission variability of point sources. E.g., consider Cusworth et al. (2021): https://pubs.acs.org/doi/10.1021/acs.estlett.1c00173
- L. 67: "prone to errors owing to the substantial human intervention required". This strikes me as a bit backwards; I would expect human intervention to produce the highest quality plume detection/delineation, just as human labeling of (for example) photos of cats and dogs produces the most accurate results. The problem is that human intervention is costly.
- L. 73-74: Varon et al. (2018) reported 15-65% additional error from uncertain wind speed.

- L. 198-201: If I understand correctly, the conversion of methane concentration to change in radiance does not account for the plume vertical distribution – i.e., the plume is first vertically integrated and then a single pressure/temperature value is used for the radiative transfer calculation. Do you expect this simplification to have a negligible effect?
- L. 219-226: I found this section hard to understand. Can you explain more clearly how a classical threshold helps with / is applied in the training procedure? Is it simply to create the ground truth plume masks for the automated plume masking task?
- L. 253-255: Suggest identifying which CNNs are encoder CNNs earlier, because it wasn't clear to me that you are referring to the CNN for binary detection and the emission rate estimator (if I'm not mistaken). And the U-Net also involves an encoder branch, so this feels a bit ambiguous.
- Fig. 2 & Fig. 5: how does the logical output of the binary plume detection model get appended to the data cube that is passed to the emission rate estimator? Is it just a uniform channel of 1's or 0's?
- Section 2.5.1: It's not clear to me what purpose this 1x1 convolution serves, can you rephrase?
- L. 297: Shouldn't this be the "encoder" part of the model, not decoder? It's encoding the input data to smaller dimensionality before applying the dense layer.
- L. 333: Can you provide some discussion of why the model might tend to overestimate concentration? Could be 1-2 sentences.
- Table 1 & general: I found myself wondering at several points whether "binary classification" refers only to the yes/no plume detection test, and/or to the U-net binary segmentation task (binary classification per pixel). Please clarify this distinction.
- Fig. 8: It's interesting that your network overestimates concentrations but underestimates emissions. If the no-plume images are really to blame for this, then they seem to have a strong effect. This could be checked by retraining the final network without no-plume images. Whether or not you do that, I feel this finding deserves a bit more discussion.
- Fig. 10: Are the retrieval fields in the left column of the figure from your network or a physics-based retrieval?
- L. 381: Are you saying that your network found 14/21 plumes? Or was that the result of a classical detection scheme?
- L. 390-391: "only one quarter of the images were deemed suitable to be analysed via clustering algorithms". On what basis?
- L. 391 & elsewhere: "clustering algorithms". It's not clear to me what you mean by this. Classical thresholding is clear, but what clustering techniques are you referring to?
- L. 393: "accurately locate" I thought those statistics were for the binary detection, but this wording would suggest they were for the U-Net segmentation. See my comment above on "Table 1 & general". Can you clarify?
- L. 396: "which is considerably lower than that obtained from classical methods". Is that true? 25% mean absolute error seems similar to what's reported in the Sherwin et al. (2022) controlled release study, for which participants used classical methods. Furthermore, the classical methods appear to have near-zero bias (Varon et al., 2018; Sherwin et al., 2022), despite significant error spread, whereas your method has 17% bias. And your 40% interquartile range also seems not so different from a ~50% error standard deviation.
- Adding to my previous comment: How does the error standard deviation (spread) of your method compare with the ~50% classical error? Or put the other way, how does your 25% mean absolute relative error compare to the same quantity for the classical methods? My impression is that your ML network is much more efficient than classical methods, but likely less accurate -- and it's not clear to me that it's more precise (except in application to multiple nearby point sources, where, as Jongaramrungruang et al. 2019 point out, using a wind-direction-independent method leads to error cancellation in the emission sum). A more careful comparison of the merits of this work compared to previous methods would be helpful.

- L. 399-410: “where the error in emission rate was greater than 50%”. But as you say, Jongaramrungruang et al. did much better than that. In addition to the possible reasons you give, could it be because they did plume detection/quantification on methane retrieval fields, whereas your model mainly relies on multi-channel radiance data? I.e., they combined physics and ML methods by processing the methane retrieval before applying their network, whereas your network is fully machine-learned. That would be an interesting finding, if true.
- General comment #2: Just a perspective to consider: I feel that the primary strength of your methodology is its potential for rapid application to increasingly large satellite datasets for methane, rather than achieving the most accurate and precise point source masks and emission rate estimates. I would expect careful human analysis to be generally superior in that respect (accuracy + precision), but perhaps not by much, and certainly highly inefficient compared to your work; clearly there aren’t enough human analysts to carefully process all the data from PRISMA, EMIT, EnMAP, Sentinel-2, Landsat, etc. That your method automates plume detection/quantification with performance comparable to human analysis, even with some low bias, is a major accomplishment.
- L. 430: Again, it’s not clear to me that your method is more successful than classical approaches in quantifying emission rates, given the combination of low bias and prediction spread you find. If I'm wrong, please just clarify the comparisons throughout the manuscript.

Technical corrections

- L. 284: section 2.4.1 --> section 2.5.1
- L. 296: section 2.4.1 --> section 2.5.1
- L. 369: section heading duplicated

References

Cusworth et al. (2021) https://pubs.acs.org/doi/10.1021/acs.estlett.1c00173

Jongaramrungruang et al. (2019) https://amt.copernicus.org/articles/12/6667/2019/

Sherwin et al. (2022) https://eartharxiv.org/repository/view/3465/