Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2022-200-RC1, 2022

Comment on amt-2022-200

Anonymous Referee #1

Referee comment on "Reconstruction of high-frequency methane atmospheric concentration peaks from measurements using metal oxide low-cost sensors" by Rodrigo Andres Rivera Martinez et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2022-200-RC1, 2022

Anthropogenic CH_4 emissions often come from sporadic and localized hotspots. This presents a challenge to emission detection and quantification, because current in situ observational networks are not dense enough to monitor point sources, whereas remote sensing observations do not provide continuous temporal coverage to capture leakage. The authors propose to tackle the challenge of sparse observations with low-cost, low-power metal oxide sensors (MOS), which can be deployed at scale. By calibrating these sensors against a high-precision cavity ring-down spectrometer (CRDS) using generated CH_4 spikes, the authors show that the accuracy and long-term stability of the sensors meet the requirements for monitoring transient CH_4 fugitive emissions. Among the calibration models examined, the authors find that a second-order polynomial model and a multilayer perceptron model achieve the best accuracy measured by RMSE. The authors then find that with a proper stratification of the data in terms of the shape of the peak, they can achieve comparable performance with 25% of the data selected for training ("parsimonious traning tests"). They also demonstrate that sensor performance degradation is not substantial within six months and that the calibration model trained on one sensor is transferable to a different sensor of the same type while maintaining sufficient accuracy. Overall, the findings represent a technological advance to address the increasing need for monitoring and quantifying CH_4 point source emissions.

While I find no major issue in this work, there is room for improvement. It seems that the parsimonious tests could explore the minimum number of peaks required for training on each cluster to achieve satisfiable performance. This information, in my opinion, would be more relevant to field operations than the somewhat vague 25% guideline presented in the work. It is also possible that the polynomial regression may need fewer cases to train than the multilayer perceptron model to achieve the same performance, and knowing how many peaks are needed for training would help inform a cost-effective strategy for deployment.

Another issue is with the cross-sensitivity. Because it is stated that MOS sensors are sensitive to electron donors other than CH_4, I wonder if the presence of ethane in natural gas would cause a problem. This potential limitation would need to be accounted for or

acknowledged at least.

I do not see a Data Availability section, and I suggest the authors check if they conform with the journal's data policy.

The writing is overall quite clear, but some grammatical errors and typos need to be fixed.

Specific comments

L10–11: "The obtained relative accuracy is higher than 10% to reconstruct the maximum amplitude of peaks (RMSE ≤ 2 ppm)" - There is ambiguity in "higher accuracy" - does it mean that the RMSE is lower than 10% of the peak amplitude? If so, it is better to say that the relative accuracy is **better** than 10%.

L24: "Anthropogenic CH_4 emissions account for 60% of global emissions (Saunois et al., 2016)" - This figure may be updated with the latest Global Methane Budget estimates (Saunois et al., 2020, ESSD, https://doi.org/10.5194/essd-12-1561-2020).

L60: "based on the observed voltage of each sensor and other variables" - What are the other variables?

L63: How would you expect to capture a spike of "several tenths of ppm" above the background (Kumar et al., 2021) using a sensor with accuracy no better than 0.8 ppm?

L101: §2.1.1 describes only five of the six chambers.

Table 1: Why is Chamber B excluded?

L132: Is there a compelling reason to downsample the data to 5 s resolution instead of 2 s?

L153: Does β represent 3.5 ppm or 3.5 standard deviations?

L273–275: This sentence seems to belong to the methods.

L288 and Fig. 5: It appears that the peaks measured by the Type E sensor lag behind those measured by the CRDS. Has the time lag been accounted for properly? Why do the peaks measured by the Type E sensor appear more dampened than those measured by the Type C sensor when both were in the same chamber?

L301: "interquartile range (IQ) = 0.001" - The interquartile ranges presented in Fig. 6 seem substantially larger than 0.001.

L303: Again, check the interquartile ranges. Unless I'm misreading Fig. 6, the interquartile ranges seem substantially larger than indicated here.

Figs. 7 and 9: Remove the axis on the right-hand side of each panel; it's unnecessary and potentially confusing. Instead, indicate that the gray dashed lines represent the target accuracy of RMSE = 2 ppm or MSD = 4 ppm^2.

L362–363: "We observed that after six months, the RMSE error produced by the models increased from 0.57 to 0.85 ppm." - This sentence is confusing. I thought the RMSE from the first experiment was 0.57 ppm for a moment, but it turned out to be the difference in RMSE. Please rewrite to clarify.

L397: "poorest" -> "poorer" - You are only comparing two sensors.

L401: Does the carbon filter create a barrier to diffusion?

L428–429: "... an RMSE of the residuals of 0.043 μmol mol^{-1} (0.69 ppm)" - This statement doesn't make sense, because μmol mol^{-1} and ppm are the same units, unless by ppm you mean something different from the volume fraction.

L439–440: "we were able to reduce the length of the training dataset from 70% to 25% while maintaining similar performance" - But the caveat here is that you need to use 70% of the data for a certain cluster and 10% of the data for all the rest of clusters to achieve optimal performance (25% of all peaks). Without a careful characterization of the diversity of spike shapes, we won't be able to know which cluster(s) to prioritize when collecting training data.