Authors' answers to RC1's comments

We thank the reviewer for his fruitful review of our manuscript. We have carefully considered all the comments and revised the manuscript accordingly. Below our answers to the commentaries raised by the reviewer.

General comments

1. **It seems that the parsimonious tests could explore the minimum number of peaks required for training on each cluster to achieve satisfiable performance. 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.**

   The parsimonious test conducted have explored the relative distribution of peaks required to have a good reconstruction of the spikes with a limited dataset. We consider that this approach will be more interesting since we explored several configurations of training sets with relative distributions of the peaks inside each configuration. For a cost effective deployment strategy, it will be necessary have a better understanding of other factors, like cross-sensitivity to other gases, before develop a more in-dept analysis of the number of spikes required to train the models.

   We acknowledge this comment. We included a statement in the conclusion section:

   "Understanding the number of spikes required for the training of the models can help to define a cost-effective strategy for deployment of the sensors."

2. **Another issue is with the cross-sensitivity. Because it is stated that MOS sensors are sensitive to electron donors other than CH4, 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.**

   The principal idea of the experiment was to characterize the capabilities of TGS sensors to measure enhancements of CH_4 similar to typical CH_4 leaks found on oil and gas facilities. The experiment was designed to expose the sensors in a controlled environment where we can simulate the spikes of CH_4 without other interfering gases. The next step following this study will be an extensive characterization of interfering gases present on gas production facilities. We acknowledge this comment, so we propose a statement in the discussion section:

   "The presence of other electron donors, such as ethane, isobutane, etc., also needs to be accounted in the model as a predictor or in the correction of the baseline."

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

   A statement was added on the manuscript.

   "The dataset was collected and the codes developed in the frame of the Chaire Indutrielle Trace ANR-17-CHIN-0004-01. They are accessible upon request to the corresponding author."

4. **The writing is overall quite clear, but some grammatical errors and typos need to be fixed.**
   The writing of the manuscript was reviewed to correct the grammatical errors and typos.

Specific comments

1. **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 <u>better</u> than 10%.**

The reviewer is right regarding the meaning of this sentence. We agree with his suggestion for the rewriting. The statement in the abstract was corrected.

2. **L24: "Anthropogenic CH4 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).**

   We have corrected the reference to Saunois et al, (2020)

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

   The sentence was corrected. Here is the new version of the sentence:

   "This study aims to compare several parametric (linear and polynomial) and non-parametric models (random forest, hybrid random forest and ANN) applied to different combinations of Figaro TGS sensors to reconstruct the CH4 signals of repeated atmospheric spikes, based on the observed voltage of each sensor. In addition, environmental variables measured from other low cost-sensors in parallel to TGS sensors, such as air temperature and pressure and H_2O mole fraction, were also added as predictors to the models."

4. **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?**

   The sentence was corrected. Here is the new version of the sentence:

   "The CH4 signal we aim to reconstruct is representative of variations observed in the atmosphere from leaks that occur within or close to an emitting industrial facility, i.e. short duration CH4 enhancements (spikes) lasting between 1 to 7 minutes and ranging from few tenth of ppm to few ppm above an atmospheric background concentration of around 2 ppm (Kumar et al., 2021)"

5. **L101: 2.1.1 describes only five of the six chambers. Table 1: Why is Chamber B excluded?**

   There is indeed an error on the sentence, we have only installed 5 chambers, but we had initially previously assembled 6 chambers. One of the chamber (Chamber B) had issues with the logging system and thus was removed from the study. We have corrected the naming of the chambers (from A to E) to prevent further confusion and the text was updated accordingly.

6. **L132: Is there a compelling reason to down sample the data to 5 s resolution instead of 2 s?**

   The noise present in the TGS voltage signal at 2s resolution is reduced with data aggregation. The choice of 5s time aggregation was a good tradeoff between the noise reduction in the signal and an adequate resolution to distinguish the peaks in the signal.

7. **L153: Does β represent 3.5 ppm or 3.5 standard deviations?**

   β is the number of standard deviations. The text was corrected.

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

   The sentence has been moved from results section (3.1 Data pre-processing and baseline correction) to the method section (2.1.2 Generation of methane spikes on top of ambient air)

9. **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?**

Yes, signals were aligned properly. It was also accounted in the correction the time lag of 10 s after applying the EWMA on the CRDS. The time lag observed on the Type E sensor is due to the carbon filter added on top of the sensing material to improve the selectivity to CH_4 on the sensor. This carbon filter also produces an airflow resistance leading to a slower response.

A statement was included on the manuscript:

"This behavior can be linked to the carbon filter included on top of the sensing material of type E sensor that produces an airflow resistance leading to a slower response."

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

    The text mentioning the interquartile ranges were corrected.

11. **L303: Again, check the interquartile ranges. Unless I'm misreading Fig. 6, the interquartile ranges seem substantially larger than indicated here.**
    The IQ ranges were corrected.

    Of note is that 2 tables have been added in the supplementary material (Table A5 and A6). These two tables show summary statistics: correlation coefficients distribution of the 20-fold cross validation.

12. **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.**

    Figures were updated following this suggestion.

13. **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.**

    The sentence was replaced by this new one:

    "We observed that after six months, the RMSE had increased in a range of 0.57-0.85 ppm on the second experiment."

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

    Accepted suggestion. The manuscript was updated.

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

    Yes, as it is stated in the "Technical Application Notes of TGS 2611 sensors" provided by the manufacturer: "the filter present on top of the sensing material of TGS 2611-E00 add an airflow resistance producing a slower response".

16. **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.**

    We acknowledge the correction. The statement was updated.

17. **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

Yes, the proposed methodology accounts for an *a priori* knowledge of the typical spikes to which the sensors will be exposed. However, in our dataset, the cluster that provides best performances in the reconstruction corresponds to spikes that covers a wide range of concentrations. This shows that models benefit from having a subset of training data containing a wide diversity of examples (here in the form of concentrations since we don't include time relationships in the models).

A statement is added in the discussion section:

"This approach is designed with an *a priori* knowledge of the typical concentrations the sensors will be exposed. Although, exposing to a wide range of concentrations, like the ones included on cluster 9 from our experiment, can lead to have a large variety of examples for the training of the calibration models."