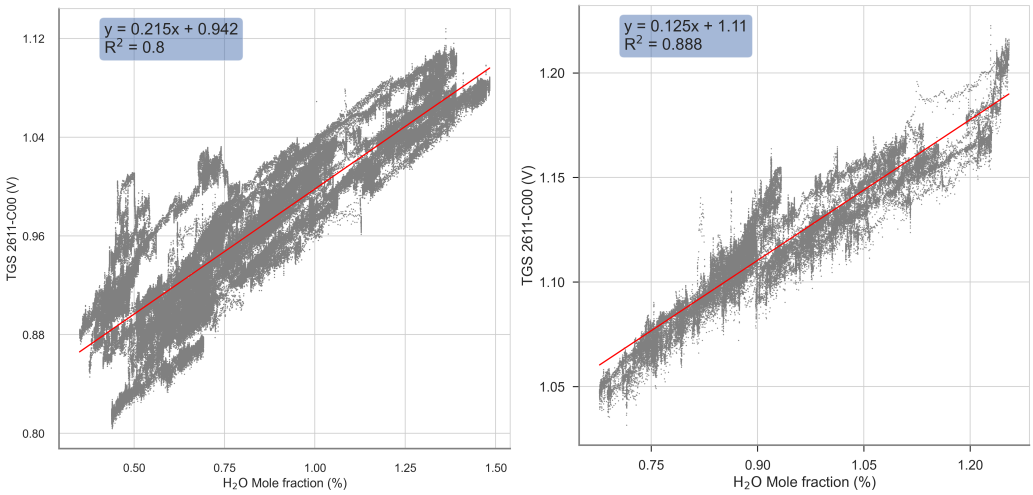
Authors' answers to RC2's comments

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

Specific comments

1. **L23-24: The statements read as if natural gas accounts for all the anthropogenic CH4**
   A statement was added:
   "Emissions from natural gas production accounts for 63% of the total emissions from fossil fuel production and use (Saunois et al., 2020)."
2. **L65-66: How could the influence of other VOC on the measurement be addressed?**
   A statement was added:
   "The main focus of this study is the behavior of TGS sensors that are exposed to enhancements of CH_4 on top of background signal without the presence of other interfering gases. The influence of VOCs on a real deployment should be considered and included as a predictor to the reconstruction models, corrected on a preprocessing stage by determining the sensitivity of TGS to them or determine, from specific laboratory experiments, the amount of signal that models can filter out and the needs in terms of ancillary measurements."
3. **L89-90: It would be worth mentioning here how the three Figaro TGS sensors differ from each other, and why only Types C and E were used for the analysis.**
   A table was included in the supplementary material and the text was updated to explain the differences between the sensors:
   "Each chamber contained a Figaro TGS 2600 originally designed to measure VOCs but sensitive to CH_4, TGS 2611-C00 with enhanced sensitivity to CH_4 and TGS 2611-E00 that includes a carbon filter on top of the sensing material to improve the selectivity to CH_4 even further (see Table A7 for information on the differences of each TGS sensor), alongside a relative humidity and temperature sensor (DHT22 or Sensirion SHT75), and a temperature and pressure sensor (Bosch BMP280, see Table 1 for details)"
4. **L101-103: Does weather during the sampling time frame have any impact on subsequent analysis? For example, Figure A4 shows the humidity and temperature ranges during the experiment. During the summer months with substantially higher water vapor mixing ratios, do the H2O vapor influences ever become nonlinear? As an aside, what causes the temperature spikes in Fig A4? Is this heating from the MOS?**
   For each experiment we characterize the effect of H_2O independently by selecting observations of ambient air without the presence of artificial spikes. Then we fit a linear model between H_2O and voltage measurements from the TGS sensor to determine the sensitivities to H_2O. The figure below shows the fit of the linear models for each experiment. We have employed a linear relationship to derive the sensitivities to H_2O on both experiments obtaining a reasonable correction of H_2O effect on the baseline of TGS. While there is possibly a non-linear relationship on periods with high humidity, our experiment doesn't allows us to have such conclusion due to our limited dataset on the second experiment (only 1 month of data at the end of summer).



Sensitivities of TGS voltage to H_2O mole fraction. The left plot shows the sensitivities for the first experiment (October 2019 – March 2020), and the right plot for the second experiment (August 2020 – September 2020). The red line is the fitted linear model.

The spikes observed in the temperature on Figure A4 are linked to fluctuations of temperature in the conditioned laboratory room. The magnitudes show the temperature inside the chamber which is affected by the heating from the MOS.

In summary, yes the periods with high humidity can produce non-linear behavior on the TGS baseline, but it is difficult to provide a general answer due to lack of available data on those periods (spring-summer).

5. **L128-129: Does the time constant differ for the TGS C and E sensors? Is this one reason for the higher phase mismatch when training the models with the Type E data?**
   No, the time constant applied to the reference instrument is the same for type C and E sensors. The phase mismatch is linked to the carbon filter included on the Type E sensor.

   We have added a sentence to clarify:
   "To determine the time constant (τ) of the buffer effect of the chambers, we applied an exponential weighted moving average (EWMA) to the CRDS data with different values of τ and compared them with the shape of the response of the TGS sensor (see Fig. A1). The time constant applied on the reference instrument to compare both TGS sensor types is the same."

6. **L144-145: As in point 4, does the H2O-voltage relationship ever become nonlinear at high enough humidity?**
   Commentary already addressed on point 4.
7. **L224-225: Will the spikes that end up being influential have any dependence on the structure in the data? In other words, is this method applicable when your dominant spike structure for a real sampling site may be unknown ahead of time?**
   Yes, in our methodology the typical clusters to which the sensors would be exposed need to be known beforehand. In the case of a real sampling site, the data used to train the models need to have examples of the typical shapes and magnitudes of spikes, therefore short sampling periods of the sensors collocated with a reference instrument would be required.

   A statement was 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."

8. **L309-311: Why does the interquartile range increase when more training data is used?**
   We attribute this behavior to two problems, the first is probably due to redundancy in the training set that condition the model coefficients to certain values and thus producing higher error on validation sets that differ much from the more common values present in the training set. The second, could be linked to the presence of few spikes on each test set of the 20-fold cross validation making that the diversity of phase errors are more represented in the summary statistics.
9. **L315-317: I think this is an important result from a policy perspective, where monitoring emission magnitudes may still be of value, even if the spike phase is not exact.**
   The reviewer is right with this remark. We have added a statement on the discussion section:
   "Models have produced a reasonable estimation of the magnitude, which is important from a policy perspective, since information of the magnitude can be of value when monitoring emission magnitudes despite the errors in reconstructing the phase of the peaks."
10. **L352-352: This sentence is unclear. The MSD is larger for Case 1 than Case 11.**
    The reviewer is right with this remark. We were comparing Case 11 and 10, the typo was corrected. Below the corrected sentence:
    "First, the smallest error did not correspond to the most parsimonious training set (Case 11) but to a larger training set (Case 10, 25% of the data)."

11. **L383-385: This result speaks to point 7 above. If your peak clustering in your training data set differs (by some threshold) from your observational set, the parsimonious training often fails to meet the acceptable RMSE. Does this imply that you would need to bring a CRDS instrument to the field site for short-term sampling? It would be worth clarifying an optimal strategy for field deployment/field calibration.**
    As mentioned in the answer on point 7, periodical short calibrations would be needed in order to update the clusters, and to achieve that the sensors would need to be collocated with a high precision instrument. Concerning the optimal strategy for field deployment, from the results of the ageing effect on the sensors, if the clustering structure is invariable, or has low variation, the models would not need to be trained for long periods of time, after six months the models achieve to meet the target requirement. If they are deployed on a site with high variability on the clustering structure, periodic re-calibrations would be needed.

    We have added two sentences on the discussion section:
    "These result shows that models would require less calibrations in environments with low variations, or invariability, on the clustering structure. But for a deployment on sites with high variability on the clustering structure, periodic re-calibrations would be needed."

Technical Corrections:

1. **Figure 1: Make image a larger so chamber set up is more easily distinguished.**
   Figure updated.
2. **Figure 1: Is chamber D mentioned in the text anywhere?**
   The diagram was corrected. Chamber D was assembled but due to logging errors the data was not usable, so it was removed from the study.
   We have renamed all the chambers from A to E in the manuscript to prevent further confusion.
3. **Figure 4: Mislabeled as Type E, not Type C**
   Label on figure corrected.
4. **Figure 9: It may be helpful to label the input data on the panels directly for ease of interpretation, if it's not too wordy.**
   We decided to keep the figure as it, since the inclusion of the input data would take a large place and probably would lead to more confusion.