

We thank the reviewer for taking the time to submit a very fair and useful review. This review has greatly helped to improve our manuscript in terms of clarity and our presentation of the key outcomes of this work. We have taken all of the comments on board and have submitted our responses below (red text) to each reviewer comment (black text).

General comments: The paper is a useful addition to the growing body of literature around these low-cost sensors. In particular, the discussion of ambient vs. synthetic air for calibration and the power-law model presented in section 3 are interesting contributions. I found the paper's narrative somewhat difficult to follow, and had difficulty keeping track of the different systems and variables (e.g. System A vs. System B; the various resistance values). Possibly some minor restructuring to center the novel findings and more explicit terminology/notation throughout would help with readability.

We are grateful to the reviewer for their kind comments. We have taken this opportunity to rename certain variables to avoid ambiguity. We have also improved the conclusion section (as suggested by the reviewer) to better emphasise the key findings of this work.

Specific comments by line number:

110-130: While these previous studies report usable results with TGS2600, they used different models and system designs, and it is unclear whether the results are generalizable or tied to specific characteristics of the research sites or experiments. In our previous study (<https://amt.copernicus.org/articles/15/5117/2022/>) we did not find TGS2600 to respond to low levels of methane in a laboratory setting. This section would be improved by mention of some of the papers' caveats; for example, Eugster and Kling (2012) note in section 3.5 an  $R^2$  of less than 0.20; Collier-Oxandale et al. found different models necessary for their different sites and note that some overfitting was observed, and so on.

We value the reviewer's advice to elaborate on the limitations of the various Figaro testing studies, discussed in the introduction. We have therefore expanded the end of section 1 to highlight specific limitations of previous studies using the TGS 2600. Yet we have remained cautious to limit the level of detail, to avoid this section becoming over-exhaustive.

It is unfortunate that we previously overlooked the interesting outcomes within Furuta *et al.* (2022), due to its recent publication. We thank the reviewer for highlighting its publication and have taken this opportunity to incorporate its valuable findings into our manuscript.

While we recognise that the reviewer did not observe strong TGS 2600 or TGS 2611-C00 methane sensitivity, we suggest that this may be due to dominant effect of variability in  $[H_2O]$  in their tests. As the reviewer suggests in their own manuscript (Furuta et al., 2022), condensation and evaporation may have occurred from their chamber walls, which we now discuss in section 5.2. In fact, we encountered a similar issue during our own chamber testing work, when we first conducted tests with humidity control turned off. For this reason, we had no choice but to use humidity control, resulting in vast variations in  $[H_2O]$  (see section 3.3), albeit about a central targeted humidity level.

172 & 184: The 5k resistor choice needs justification. As you note at line 165, the reference resistor should be close in value to the expected sensor resistance. Around background methane levels one would expect a sensor resistance an order of magnitude larger than your

5k choice, as can be seen in your supplemental information (which is consistent with our observations).

The reviewer is right to suggest that a load resistor with a higher resistance would have been more appropriate in this work, based on typical resistance measurements made by the Figaro sensors. However, we chose to use the same load resistance in both System A and System B, to make measurements and systematic errors from both loggers as comparable as possible. Thus, we were effectively bound by the resistance selected by the System A manufacturer.

Therefore, we now clarify that the System A load resistor was hard-wired into the logging box in section 2.2. We now also state that we chose the load resistance in System B to mirror System A, in section 2.3.

The choice of a 5 k $\Omega$  load resistor used by the System A manufacturer can be justified by looking to load resistors used in previous work. For example, Jørgensen *et al.*, 2020 used a 10 k $\Omega$  resistor. We now make this additional point in the manuscript in section 2.2. We also include a short discussion on this point in section 4.2, where measurements from the Amailloux landfill site are presented. Although a higher load resistance would improve sensitivity, 5 k $\Omega$  is sufficient to detect resistance variations in response to changes in the surrounding environment.

169-179: More detail about the logging system would be useful, particularly the ADC resolution, noise floor, and so on, as these can be expected to determine sensitivity in combination with the reference resistor choice. I would be interested in a brief sensitivity calculation using the ADC resolution/noise floor and the 5k reference resistor to show the ability to detect small changes in the sensor resistance. The logger's power supply stability is also a critical detail for your system's performance.

This is a useful suggestion. We have provided additional details on the ADC resolution in section 2.3. We now specify that although it has a maximum resolution of 18 bit, we set it to sample at 16 bit (for improved sampling frequency), resulting in 0.15 mV resolution. This translates into a resistance precision of 0.6  $\Omega$ , assuming 5 k $\Omega$  Figaro resistance.

According to the manufacturer data sheet, the System B power supply has rated ripple and noise effects of less than  $\pm 1$  mV. In reality, we observed a standard deviation of  $\pm 0.1$  mV, when tested with the ADC. We now also provide this additional detail in section 2.3.

We thank the reviewer for the suggestion to include these details, which we previously overlooked in terms of importance. We hope that these useful resolution values support our measurement setup.

194: For the application 35mV is possibly not good enough; is this value referring to the accuracy of the setting, ripple, drift, or a combination of all of them? 35mV ripple would swamp any sensor response to small methane changes, for example. You have some discussion of supply voltage sensitivity in your supplemental information, but a short quantitative discussion of what this supply voltage tolerance means in terms of sensor response/detection capability would be illuminating, as it's not immediately obvious to me how big of a concern the supply voltage accuracy is for your experiment.

We agree that an actual 35 mV power supply accuracy (including ripple effects) would be insufficient in this work. In reality, this 35 mV value refers to a read-back setting accuracy, which we now clarify in the manuscript in section 2.3. As we adjusted the supply voltage using an independent measurement of potential difference across the Figaro circuit board, the power supply voltage accuracy has no effect on our testing. To further augment our confidence in the stability of our power supply unit, we measured the circuit board potential difference on many occasions on different days, finding no change in voltage. Nevertheless, the reviewer raises an important point here which we hope we have now resolved in the manuscript in section 2.3.

311-318: The H<sub>2</sub>O fluctuations within each period are substantial (from the chart, up to 0.5%) and appear to continue over the whole of each sampling period. If the sensors require hours to stabilize at a given humidity level, will they stabilize at all given this large fluctuation?

The reviewer makes a valuable observation regarding the large fluctuations in [H<sub>2</sub>O] during our environmental chamber tests. Our work shows that the key issue associated with unstable water, is the associated lengthy Figaro stabilisation time, whereby resistance gradually exponentially decays towards a stable resistance (see Appendix B). We showed that this resistance decay occurs immediately following a step-change in [H<sub>2</sub>O].

In the case of the environmental chamber, rather than a single step transition in [H<sub>2</sub>O], there is periodic variability about a central point. This phenomenon occurred because the humidity control system constantly works to rectify humidity towards a target setting. As this water variability occurred both above and below the target [H<sub>2</sub>O] setting, the water resistance delay effect effectively occurred in both directions simultaneously, thus cancelling itself out.

This can be observed in the top pane of Fig. 7, where after the initial temperature change, Figaro resistance appears to stabilise within 8 hours. Thus, resistance delay effects due to changes in [H<sub>2</sub>O] cancel each other out. This argument is now explicitly clarified in the manuscript, where suggested by the reviewer in section 3.3.

316 and throughout: Could you remind the reader what R<sub>2</sub> indicates or use a more descriptive subscript? I had difficulty remembering the resistance notation, much of which is similar - R<sub>2</sub>, R<sub>b</sub>, R<sub>l</sub>, etc.

To reduce ambiguity,  $R_l$  has been renamed to  $R_{load}$ ,  $R_b$  has been renamed to  $R_{baseline}$  and  $R_2$  has been renamed to  $R_{2\text{ ppm}}$ . We hope that these changes make these different resistance terms easier to follow.

337: As the sensors are much more responsive to humidity than to methane, is 3% uncertainty good enough?

This 3% uncertainty value represents the uncertainty in the  $R_{2\text{ ppm}}$  resistance baseline (at 2 ppm [CH<sub>4</sub>]).  $R_{2\text{ ppm}}$  is calculated as a first step, before subsequently calculating [CH<sub>4</sub>] enhancements above the 2 ppm [CH<sub>4</sub>] level. As  $R_{2\text{ ppm}}$  and [CH<sub>4</sub>] are calculated separately, and as different steps, it is difficult to compare the two uncertainties and to aggregate them. Therefore, it is not straightforward to make a direct link between uncertainty in  $R_{2\text{ ppm}}$  and a theoretical resulting uncertainty in [CH<sub>4</sub>]. Each uncertainty for each model must be treated independently. We now emphasise the significance of this 3% value more clearly in the manuscript in section 3.3, to avoid confusion.

As we were unable to derive  $[\text{CH}_4]$  estimates during field deployment, it is difficult to know whether this 3% uncertainty in  $R_{2 \text{ ppm}}$  had a significant influence on the overall  $[\text{CH}_4]$  uncertainty.

355: Fig 10 shows some fluctuation in sensor response in the last two minutes, and it looks like the sensors stabilize at the new methane levels quite quickly. Why is it better to select the last two minutes of each methane level rather than the last 10 minutes, which appear to already be stable?

Our original Figure 10 was used to highlight the general stability of our logging system, rather than the stability of individual Figaro sensors. This figure did not therefore well-illustrate Figaro stabilisation issues, as LSCE001 (shown in the original figure) is one of the more stable sensors that we tested. On the other hand, LSCE009 took longer to stabilise in response to changes in  $[\text{CH}_4]$ . We have therefore updated Figure 10 with an example of a LSCE009 methane transition, which shows a longer stabilisation period.

Based on the delayed stabilisation effect illustrated in this figure, we used a 2-minute averaging period for all five sensors, for consistency. Furthermore, Figure 11 justifies our choice of averaging time, as the  $[\text{CH}_4]$  points fit the model curve very well. This satisfies the ultimate aim of this test to characterise methane sensor response. In theory, taking the final minute instead of the final two minutes, would probably produce the same result with no advantage either way. We have added a sentence to section 3.4, to support this reasoning.

361: Why is Eq. 3 only valid for system A?

We thank the reviewer for raising the ambiguity in this statement. In fact, Eq. 3 is an entirely empirical logger-specific model used to relate measured temperature and derived  $[\text{H}_2\text{O}]$  to measured resistance. However, Figaro resistance is also influenced by logger-specific parameters such as airflow and thus associated cooling effects. Furthermore, the gradient between the point of each temperature measurement and each Figaro sensor is also logger-specific, especially for System B, where five different Figaro sensors were tested in the same logging cell. Therefore, Eq. 3 model parameters can only be used for the logger in which they have been derived. We now clarify these points in section 3.4 and section 3.3.

Table 4: The variation in the alpha values for the sensors is surprising to me - our previous work found TGS2611-E00 to be quite consistent, at least within the same production batch. Were your sensors taken from the same batch, or is there some other component in the system that might be causing this variation? You mention this at line 641, but it would be good to also indicate whether your sensors have the same or different batch codes (printed on the side of the component).

Unfortunately, it is difficult to satisfy this point directly at this moment, as the sensors are currently deployed in the field. However, we know that LSCE001, LSCE003 and LSCE005 were purchased by Scientific Aviation in the USA, whereas LSCE007 and LSCE009 were purchased by us at a later date in France. So, we are fairly certain that at least two different batches were tested, but there may have been more. We now add the point that the sensors likely come from at least two batches in section 5.2 of the manuscript. Unfortunately, it is difficult for us to elaborate any more on this without certainty on specific sensor batches.

421: Again, why the last five minutes? It looks like the sensors stabilize more quickly than that, as far as I can see from Fig. 12.

The averaging duration was chosen as a compromise between maximal stability and maximal averaging points. As the sensors stabilised much faster in this carbon monoxide test compared to the methane test (see response above), a longer averaging period of 5 minutes was more appropriate here, than 2 minutes used for the methane test.

We now highlight this point in section 3.5 of the manuscript and explain our rationale for selecting a 5-minute averaging period. In any case, we do not feel that prolonging the averaging time would have a significant impact on the outcomes of this test: choosing a 10-minute averaging time instead of a 5-minute averaging time would probably yield the same qualitative conclusion.

Section 5.1: Your field and lab tests presumably used different power arrangements. Could you add some discussion of the steps taken to ensure consistent electrical operating conditions between the field and lab tests (particularly 5V supply stability)? In the supplemental information you show that different supply voltages cause different sensor responses; is this possibly involved in the differences?

The reviewer makes an interesting point which we previously failed to clarify. The System A logging system converts battery voltage into a stable 5 V Figaro power supply over a wide battery supply voltage range. Therefore, despite the 12 V lithium ion phosphate battery being connected to a charger instead of a solar panel, the Figaro power supply remained unaffected. We now make this point clear in section. 5.1 and also in section. 2.2, where we introduce the System A field logger.

Section 6: This section is difficult to read to me, and doesn't highlight the major contributions of the paper. It would be more clear to me if broken into multiple paragraphs, and with stronger emphasis on the findings you believe to be particularly important.

We appreciate this useful suggestion and have broken section 6 up into smaller paragraphs. We have rephrased the existing points in this section to improve clarity. We have also included additional key points, to emphasise the key findings of our work, as suggested by the reviewer. For example, we now stress the value of the laboratory testing results, where we characterised methane response using an adapted power fit.