Reviewer 1:

Overall, this study presents an interesting and well-written technical paper, showcasing a low-cost alternative to conventional CO_2 measuring devices. However, I miss a few critical details, that I think are needed to make it publishable.

First, I think a more detailed evaluation against LI-COR data is needed. Currently, there are few details, and a more in-depth analysis of discrepancies, especially on systematic biases, would be needed. Additionally, it would be good to see if there is temporal drift and to examine in more detail why there are certain times at which there is a relatively large mismatch.

Further, I missed a bit the discussion on how these methods could be applied in low-cost settings. E.g., how accurate are they, if no calibration data from a more expensive system is available? And how cheap are they really, if accounting for the time invested? I think it would be good to give an estimate of the time to assemble it and get it running from scratch. For people wanting to try this, these "time costs" may be very relevant.

Finally, I think it would be important to stress the limitations more clearly. Especially the geographic/climatic scope should be well described. For example, I could imagine that the device would work much worse in a more humid environment, if CO_2 evolution after rainfall is a major issue (as you seem to indicate).

We want to thank the reviewer for the positive review and the constructive comments, which helped us improve the manuscript. Detailed answers are given below using blue font. We would like to emphasize the main changes:

(1) We added a more detailed evaluation to compare the modeled gradient flux with the LI-COR chamber flux. Specifically:
+ An assessment of sensor drift over time.
+ Additional statistics (RMSE, components of mean squared error) for the evaluation of modeled gradient fluxes using measured chamber flux.
+ Additional interpretation for the fitting performance of diffusion models.
+ Discussion on the discrepancies between the best-fitted flux and the measured flux on rewetting events.

(2) We added a discussion on the application of the system in low-cost settings in Section "3.3 Limitations and modifications". Specifically:

+ Suggested alternative low-cost validation methods to the LI-COR chamber flux system.

+ Time cost evaluation is added to the do-it-yourself guide.

+ The geographic/climatic scope for using the device and the gradient method.

First, I think a more detailed evaluation against LI-COR data is needed. Currently, there are few details, and a more in-depth analysis of discrepancies, especially on systematic biases, would be needed. Additionally, it would be good to see if there is temporal drift and to examine in more detail why there are certain times at which there is a relatively large mismatch.

We address the three key points in the above comment:

- First, we analyzed raw data to determine if there was a gradual drift over time and if there was a need for more than one calibration curve for each sensor. For that, we calculated pairwise concentration differences (sensor#2-sensor#1, sensor#3-sensor#1, sensor#2-sensor#3, 3 pairs/depth, 2 depths 5 and 10 cm). The results were added to Figure S4 in the supplementary information (attached below).

  We calculated the relative deviation of each sensor/pair, multiplied by the total number of days, and then compared it to the average concentration. If the relative deviation <= 10% of the average concentration, one calibration curve is needed. Otherwise, separate calibration curves for each period where the relative deviation remains below 10% are required. It is noted that the 10% is the predefined threshold used in this study; depending on the accuracy standard, this threshold can be adopted differently.

  An example of relative deviation calculation:

  For sensors at 5 cm:

  Sensor #2-#1: relative deviation = -0.08*175 = -14 ppm

  Sensor #3-#1: relative deviation = 0.14*175 = 24.5 ppm

  Sensor #2-#3: relative deviation = -0.22*175 = -38.5 ppm

The relative deviations are relatively small compared to the concentration range at 5 cm (300~650) and are also lower than 10% of the average CO_2 concentration. Therefore, one calibration curve/sensor was used. We added this analysis to the manuscript with a reference to Figure S4 below.

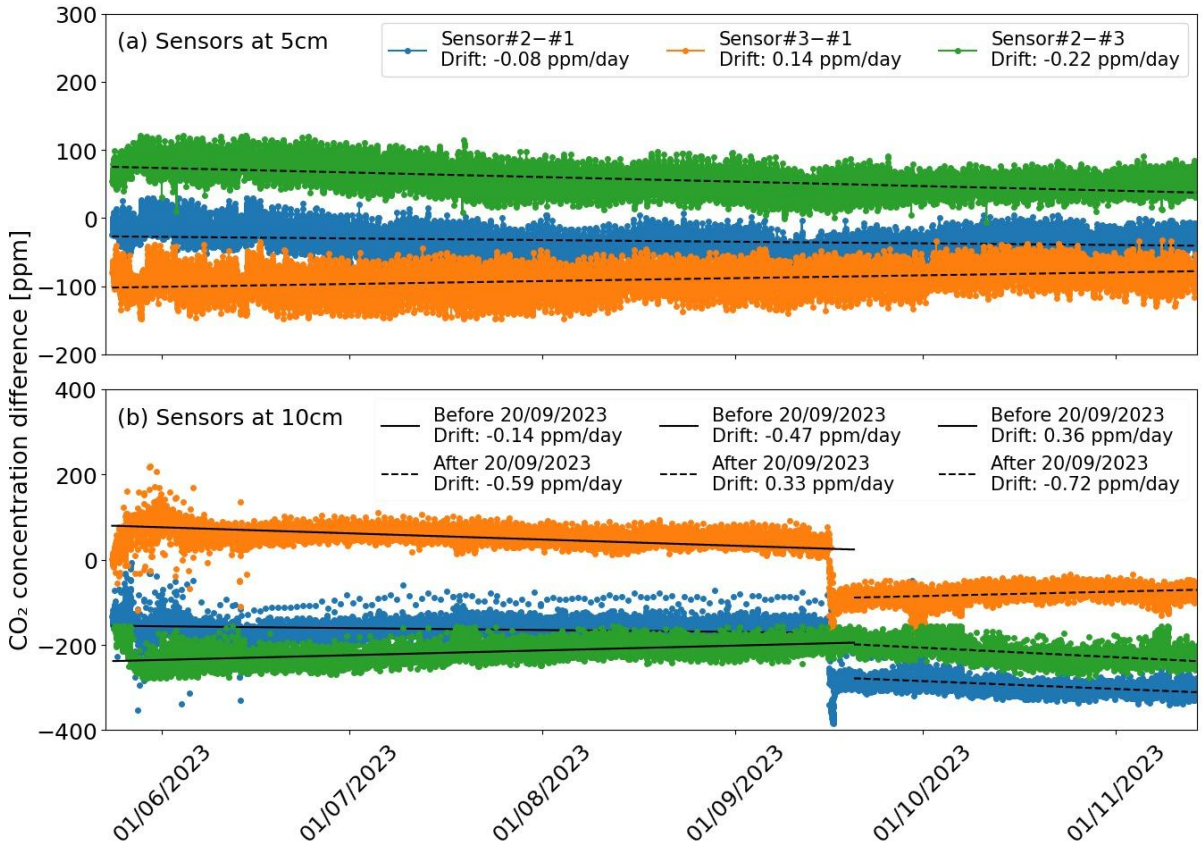

Figure S4. Pair-wise CO_2 concentration differences between three sensors at 5 cm (a) and three sensors at 10 cm (b). The drift rate for sensors at 10 cm was evaluated separately for two periods, before and after 20/09/2023, when the baseline of sensor#1_10cm drifted systematically from ~300 to ~200 ppm.

- Second, we added more statistics for the comparison and the selection of the best diffusion model. Specifically, we added components of mean squared deviation: squared bias, non-unity slope, lack of correlation (Gauch et al., 2003), and RMSE. The mean squared deviation analysis is presented in Figure S6 in the supplementary information with a relevant reference in the manuscript.

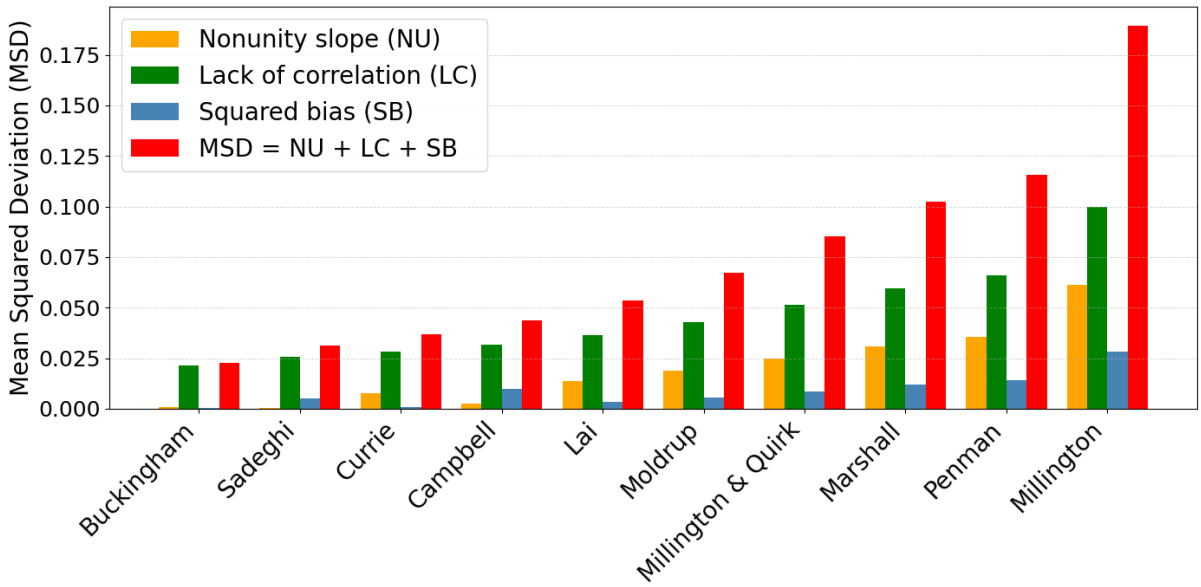

Figure S6. Components of mean squared deviation (MSD) for ten diffusion models. The three components are lack of correlation (LC), non-unity slope (NU), and squared bias (SB). Data for comparison are calculated gradient fluxes (F_{GM}) using ten diffusion coefficient models and chamber flux (F_{CM}) measured by the LI-COR chamber and gas analyzer.

From the boxplot analysis (new Figure 5a, attached below) and mean squared deviation analysis (Figure S6, attached above), the Buckingham diffusion model showed the best performance. Therefore, it is selected for further analysis to see, on a daily or hourly basis, where the mismatches occurred.

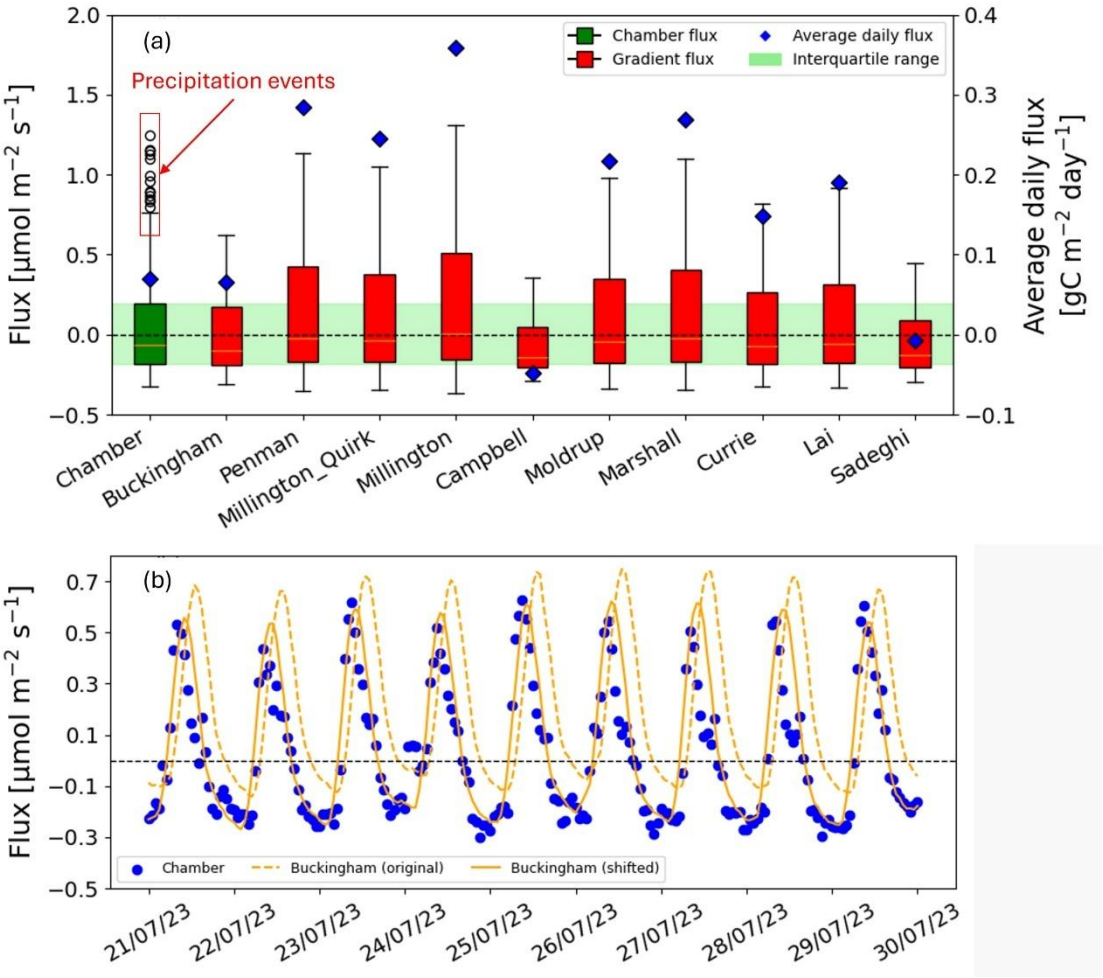

Figure 5: Comparison of measured chamber flux (green) and calculated gradient flux (red) using ten published gas diffusion models, and average daily cumulative flux (blue scatter) (a). Diurnal cycles of measured chamber flux (blue scatters) and calculated gradient flux using Buckingham diffusion model (dashed orange) and Buckingham gradient flux shifted by 3-hour lag time (solid orange) during nine representative days without precipitation (b).

- The large mismatches observed when comparing Buckingham modeled gradient flux and measured chamber flux occurred during rewetting events. This is a known methodological limitation of the gradient method and not the focus of this study. We extended the discussion to make this issue clearer.

"The observed CO_2 pulse, as measured by the CM, agrees with the observed pattern of very high rates right after rewetting and slowly declines over time (Kim et al., 2012). These precipitation-induced CO_2 pulses were underestimated by the GM. Previous studies also

reported that the GM did not capture the abrupt CO_2 pulse increases after water application (Jiang et al., 2022; Yang et al., 2018). Rewetting of arid soils after a dry period triggers the sudden increase of microbial activity, leading to a burst in carbon mineralization (Barnard et al., 2020). In arid soil, the top ~1 cm is often the most microbially active due to the presence of biocrust (Weber et al., 2016). The increased CO_2 efflux from the topsoil was captured by the CM, yet underestimated by the GM (Jiang et al., 2022; Yang et al., 2018). Under rewetting events, the assumptions of the GM, such as one-directional gas movement and linear concentration gradient with soil depth, are invalid. Greater soil CO_2 on the topsoil than in the deeper soil leads to bidirectional concentration gradients and fluxes (Tang et al., 2005). The application of the GM, therefore, is not recommended for F_S estimation of dry soils upon rewetting. It is important to note that this is a well-known methodological limitation, extensively reported in the literature, and it persists regardless of the type of NDIR CO_2 sensor used (Fan & Jones, 2014; Tang et al., 2005). Even though F_{GM} under rewetting events is unreliable, it does not limit the application of the GM under relatively steady moisture conditions (i.e., SWC can be moderate to high but not change due to rainfall or irrigation) (Fan & Jones, 2014; Turcu et al., 2005)."

Further, I missed a bit the discussion on how these methods could be applied in low-cost settings. E.g., how accurate are they, if no calibration data from a more expensive system is available? And how cheap are they really, if accounting for the time invested? I think it would be good to give an estimate of the time to assemble it and get it running from scratch. For people wanting to try this, these "time costs" may be very relevant.

- For the application of the low-cost sensor system in low-cost settings, we suggested more methods for validation of the modeled gradient flux.

"Several alternatives can be considered. First, the site-specific diffusion coefficient can be measured directly for the calculation of F_{GM} without using published gas diffusion models. For example, Osterholt et al. (2022) suggested an approach to inject CO_2 as a tracer gas to estimate the diffusion coefficient. Furthermore, high-end, expensive chambers and gas analyzers can also be replaced with a low-cost, open-source chamber system (e.g., Forbes et al., 2023). The same CO_2 sensor SCD30, as used in this study, can also be used to manually build a low-cost chamber. When used with the LC-SS, only one chamber-gas analyzer system per several LC-SSs is needed since

only a short duration of F_{CM} measurements is required for validation. Additionally, conventional CO_2 quantification techniques - such as gas chromatography or the alkali absorption method - can be used to monitor CO_2 concentration changes inside a static chamber to quantify F_S (Yan et al., 2021; Pumpamen et al., 2004; Yim et al., 2002; Christiansen et al., 2015). Integrating the LC-SS with the alkali absorption method could be a promising approach that balances affordability, automation, and long-term monitoring of CO_2 concentration and F_S, while enhancing accuracy; particularly in remote or resource-limited locations where access to high-end instruments like gas analyzers or gas chromatography is not feasible."

- On the accuracy of sensor systems for quantifying CO_2 concentration, even for high-end sensors such as the commonly-used Vaisala sensors (https://www.vaisala.com/en/products/instruments-sensors-and-other-measurement-devices/instruments-industrial-measurements/gmp252), field calibration is almost always a must for the insurance of data quality. Therefore, we do not recommend using the raw data without calibration (as mentioned in the manuscript).
- Our low-cost system, ~ 700 USD/per system (with an additional building time of about ~3 days/per system), requires around 1 day every few months for CO_2 sensor calibration and flux validation. In our opinion, this can still be considered low-cost. In fact, we have been running five copies of this system along the Negev desert for the last year. This showcases the potential significant opportunities for these types of systems in large-scale, long-term comparative research. In any case, we added the building time to our GitHub DIY guide.

Finally, I think it would be important to stress the limitations more clearly. Especially the geographic/climatic scope should be well described. For example, I could imagine that the device would work much worse in a more humid environment, if CO_2 evolution after rainfall is a major issue (as you seem to indicate).

We added to the limitations section the geographic/climatic scope where we think the system will not work well.

"The second limitation is that the system was tested only in dry, arid soils. Although a few precipitation events were captured and analyzed, the system's performance under persistently high SWC conditions was not evaluated over the long term. In general, the use of the GM may not be

suitable under conditions of sustained soil saturation, frequent rainfall typical of humid climates, or frequent irrigation."

Detailed comments:

L46 it would be good to show some evaluation stats for your comparison to LI-COR.

We added the RMSE value (0.15 μmol m^{-2} s^{-1}) to the manuscript.

L59 You might also mention the importance of Fs data for (agro)ecosystem and soil carbon models in calibration, validation, and development.

We added the above insight and citations to the manuscript.

L145 Since you are talking about CO_2 flux, it would be better to give soil organic carbon, not soil organic matter (usually SOM/1.72).

We measured soil organic carbon (9.37 mg/g and 9.13 mg/g for soil at depths 5 and 10cm, respectively). We added the values to the manuscript.

Section 2.5. I think you should do more than just a linear regression with LICOR data to check for the accuracy of your approach. For example, display of relative RMSE, analysis of whether there is a systematic bias, and a slope of the regression different from zero (e.g., Gauch et. Al., 2003, https://www.agronomy.org/publications/aj/abstracts/95/6/1442) are needed for proof of good performance. Also, it may be important to dissect where the systematic differences between sensors in Figure 3b (raw data) stem from.

For the validation of modeled gradient fluxes using measured flux (Section "2.5. Validation of F_{GM} using F_{CM}"), besides the comparison using boxplots (Fig. 5a) and the linear regression (Fig. 6a), we added the mean squared deviation components as suggested in Gauch et. al, (2003) (added as Figure S6 in the supplementary) and the RMSE between the measured and the gradient flux modeled by Buckingham equation (added to Fig. 6a) (both figures are attached above). In addition, we added the suggested reference to the manuscript.

The systematic differences between sensors in Fig. 3b raw data occurred most likely because of the automatic baseline correction (ABC) algorithm (as a default algorithm for drift correction in

open-air environmental settings). In soil, since sensors were not exposed to fresh air, so the ABC miscalculated and adjusted each sensor's baseline differently.

L229 I disagree that your displayed results validate the stability after 6 months. They only show how much correction was needed but not if there was a temporal trend in the correction needed (e.g., drift of one sensor from the other two). The latter would be interesting to analyze in detail. In the simplest terms, you could do this by plotting the sensor's raw data over time. Maybe calculate correlation statistics for each month individually.

We agree that more analysis, such as a temporal trend of gradual drift, is necessary to show the system's stability after six months. As mentioned above, we analyzed the relative gradual drift of one sensor compared to the other two at the same depth. The results are shown in Figure S4 in the supplementary section and added to the discussion in section "3.1. CO_2 sensor calibration".

"Gradual drift was assessed by evaluating whether the pairwise differences in CO_2 concentration among three sensors placed at the same depth (sensor#2-sensor#1, sensor#3-sensor#1, sensor#2-sensor#3) changed over time. To quantify this, the pairwise concentration differences were plotted against time, and linear regression was applied to determine the relative drift rate (ppm day^{-1}). The cumulative deviation was then estimated as the product of the drift rate and the number of days. If this cumulative deviation exceeded a predefined threshold - set at 10% of the mean concentration in our study - separate calibration curves were applied to account for the drift."

"Over the tested period, we observed a low rate of gradual drift in all six sensors (0.06-0.72 ppm day^{-1}) (Fig. S4). The cumulative deviations for six sensors were below the predefined threshold - 10% of the mean concentration. Therefore, for the entire period of 175 days, we used one calibration curve for each sensor."

Figure 3: More information is needed in the caption so that this figure can stand on its own. E.g., over what time periods were reference and SCD30 measurements made. Additionally, it would be nice to code Fig 3 a by date of the measurements made.

We changed the caption of Figure 3 as suggested to make it stand-alone.

"**Figure 3**: Calibration curves of the SCD30 CO_2 sensors using reference CO_2 concentration measured by Vaisala CO_2 sensor between 12/6-17/7/2023 and LI-COR gas analyzer 10-11/9/2023

(a), and distribution of CO_2 concentrations collected by six SCD30 CO_2 sensors after field and statistical calibration (b)."

Figure 4: also here, more information is needed so that the figure can stand on its own. E.g., abbreviations are not defined (Fs, Tsoil, SM). Panels c, d, and e should be amended with the exact dates that they refer to. Or are these averages? This is not clear for me.

We changed the caption of Figure 4 as suggested to make it stand-alone. We also added text and error bars to Fig. 4c, d &e to clarify that average data are presented.

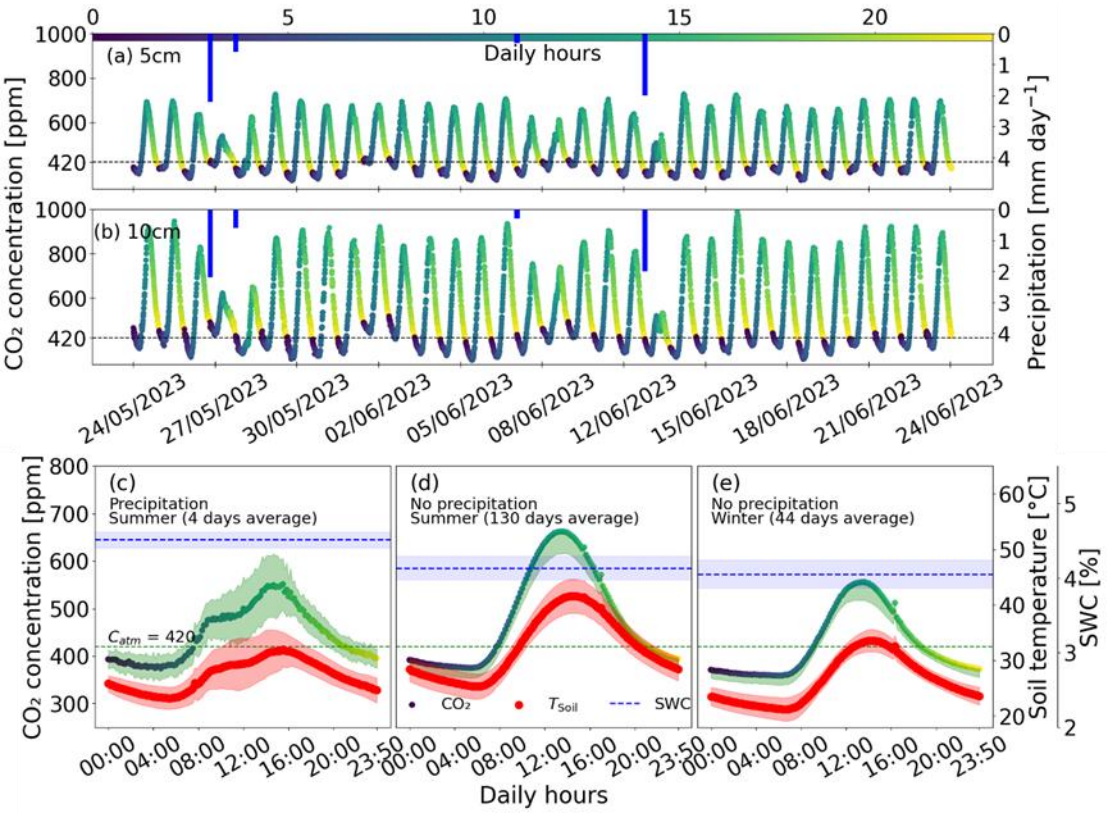

Figure 4: One month example of continuous CO_2 concentration measurements between 24/05-24/06/2023 at 5 cm (a) and 10 cm (b) depths, average daily values at 5 cm of CO_2 concentration, temperature, and volumetric soil water content (SWC) during four days with precipitation from May to September (Summer) (c), 130 days without precipitation between May and September (Summer) (d), and 44 days without precipitation between October and November (Winter) (e).

L290 to 291 This part belongs to the Methods section.

This part was moved to section "2.4. Calculating the F_{GM} using the LC-SS data".

Figure 5 a) is quite messy with all the lines and hard to read. I suggest you move this into the supplement and show only the best-fitting method, here. Again, make sure to define all abbreviations so that the figure stands on its own.

We replaced Fig.5a and kept only the best-fitted Buckingham method. A figure with all predicted fluxes is presented in Figure S5 in the supplementary. The caption was changed as suggested to make it stand-alone.

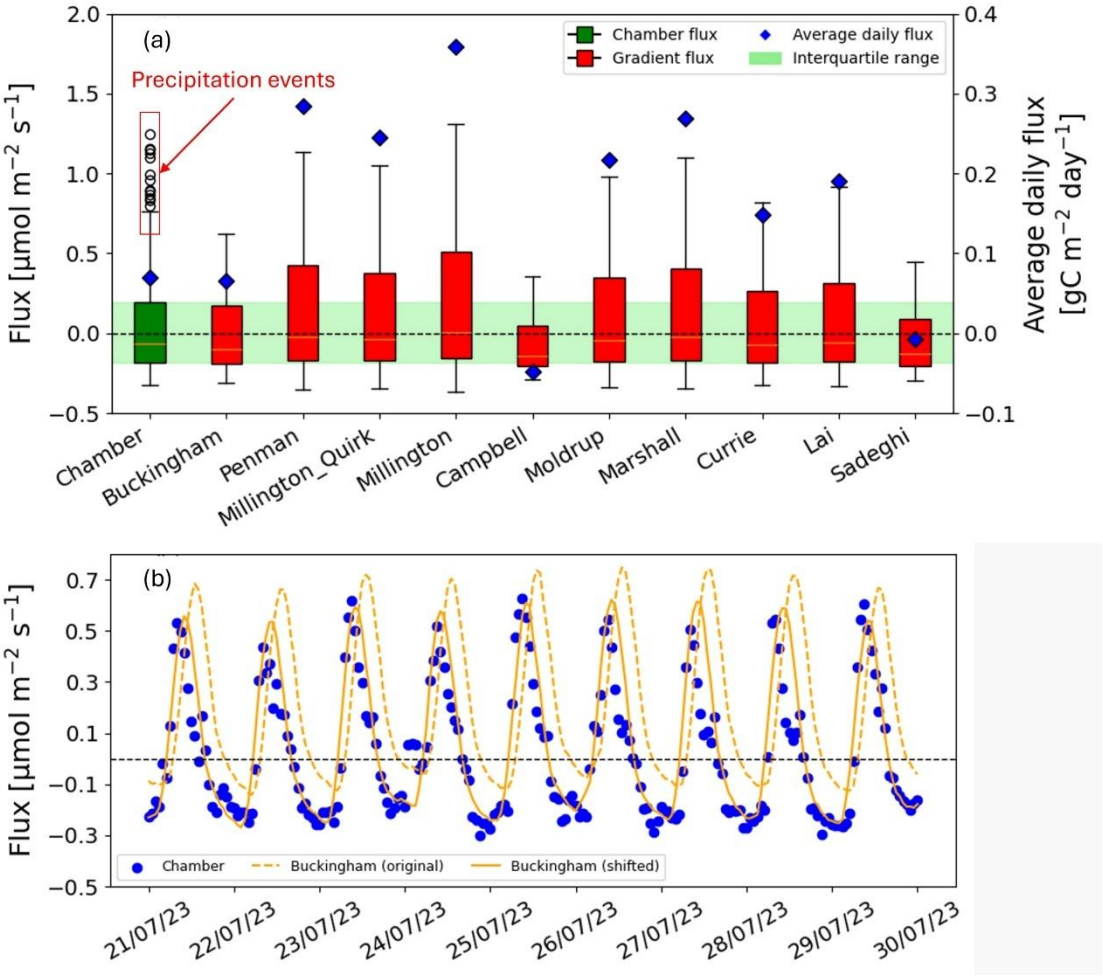

Figure 5: Comparison of measured chamber flux (green) and calculated gradient flux (red) using ten published gas diffusion models, and average daily cumulative flux (blue scatter) (a), and diurnal cycles of measured chamber flux (blue scatters) and calculated gradient flux using Buckingham diffusion model (dashed orange) and Buckingham gradient flux shifted by 3-hour lag time (solid orange) during nine representative days without precipitation (b).

L306 do you mean “correlated most strongly”?

Yes, the sentence was corrected to “correlated most strongly”.

Figure 6. Please also report additional evaluation statistics, as suggested above. And again, please define all abbreviations to make the figure stand on its own.

We added RMSE as additional evaluation statistics to Fig. 6a and changed the caption to make it stand-alone.

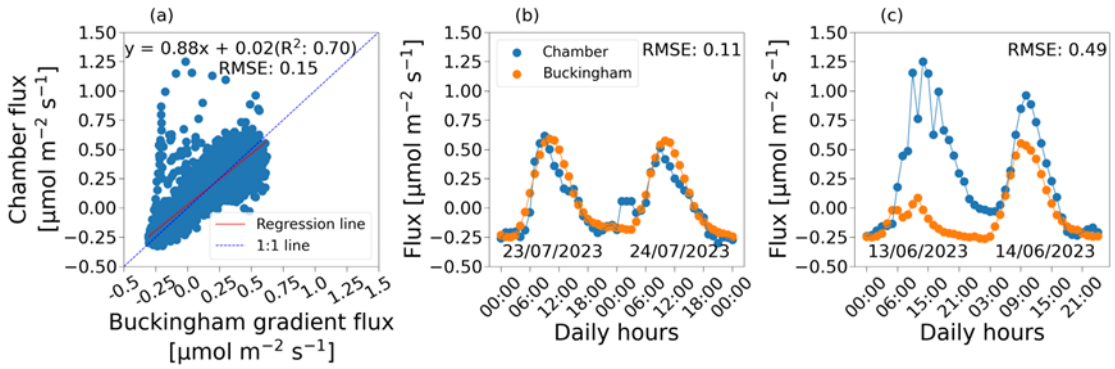

Figure 6: Comparison between the gradient flux (F_{GM}) calculated by the best-fitted Buckingham diffusion model and the LI-COR chamber flux (F_{CM}) for the whole tested period of 175 days (a), the Buckingham gradient flux (orange) and the LI-COR chamber flux (blue) during two representative days without precipitation (23-24/07/2023) (b), and during two representative days with precipitation (13-14/06/2023) (c).

Section 3.3 is a bit short and I missed a clear recommendation where your system is to be used and where it could have limitations. For example, does it only work well in arid regions (your mismatch mainly occurring on rainy days could suggest this)? Additionally, what methods could be used to improve the reliability of your system? What cheaper methods to double-check your results could you recommend? Could your system even be applied in a manually built chamber to conduct the chamber method?

As mentioned above, we added to the manuscript alternative methods, such as the alkali absorption method, as a cheaper and more accessible method to double-check the modeled gradient flux and also added more discussion on the limitations and a clearer recommendation of where and when the system may/may not work well.

L345 this is the first time you mention maintenance requirements. It should be in the results/discussion if you want to have it in the conclusion.

We removed it from the conclusions section.

L350 I think you should still mention the limitations and potential geographic/climatic suitable ranges (i.e., you only tested it in a very arid environment).

We fully agree – please see our previous answers above.