Response to Anonymous Referee #2

Responses in blue

General comments

This study characterizes the performance (accuracy, precision, drift) of one type of low-cost CO2 sensor for ambient air measurements. The paper describes an experiment where duplicate sensors are used to make continuous ambient measurements in an environment with conditions that change slowly over time. Ambient pressure, temperature, and humidity are monitored simultaneously and used to derive empirical corrections for individual sensors, which significantly improve the accuracy of the final datasets. The overall experiment is well-designed and the analysis of the resulting data is sound.

The paper topic is highly relevant and potentially useful to the broader atmospheric measurement community, but currently falls short of that potential. There are several additional experiments and analyses one could imagine that would fit in the same paper and would improve the scope and significance, such as: (i) test whether a unique correction is needed for each unit and what the uncertainty would be if a generalized correction were applied, (ii) demonstrate an experiment which would allow correction factors to be rapidly derived in the lab, (iii) test the K30 in a real-world environment, (iv) test the K30 for long-term drift. At a minimum, it appears that the authors can use the existing dataset to address point (i).

First off, thank you very much for your thorough review of our paper and for the helpful comments and suggestions. For your general comments:

(i) This additional analysis has been added to Section 6 as was suggested, but to summarize, yes a unique correction is required. By taking the average coefficients and intercepts computed by the multivariate regression of the five best performing K30s, all six sensors had higher RMSEs than with independent coefficients, which is expected. But at a minimum, the RMSE doubled with in some cases actually became worse than with no correction at all. Thus, instead of 1.5-3 ppm RMSE after correction, it can range from 3 to 24 ppm for 1 minute data.

(ii) The main reason why we did an ambient calibration, and the main concern with doing a laboratory correction is the labor and equipment required to do this. To cycle temperature, pressure, and relative humidity throughout typical ambient ranges is not difficult for one instrument, but for several requires a large enough chamber, and needs to be set up to be autonomous, otherwise if someone is manually controlling these parameters for days/weeks, the low-cost aspect of these instruments becomes much more labor intensive. At this time, we did not have the resources or an environmental chamber available to conduct this experiment. This is something we hope to do with future work, as mentioned in the conclusions.

(iii) The experiment described in the manuscript is a quasi-real-world environment, as we used ambient air and ambient variations in the environmental variables, but the main difference was the sensors were not outside in direct sunlight or exposed to weather. This is something we hope to do in the future, but we wanted to evaluate them without these engineering concerns initially.

(iv) This is also something we hope to achieve with future work (as mentioned in the conclusion portion of the paper). While it would have been nice to include long-term drift in this manuscript, that would require 6-12 months of data with either a continuous gas analyzer as reference or calibration gas introduced periodically. We believe it is sufficient to publish these initial results characterizing month-long drift, as the results will be useful to others working with similar sensors. We will include an evaluation of long-term drift of this sensor and others in a future manuscript.

Specific comments

Los Gatos Instrument

You use the LGR dataset as a control, but I am concerned that there could be large uncertainty associated with the LGR water correction. Do you know the accuracy of the LGR water correction? I have never seen an assessment of it.

Yang et al. 2016 describes a comparison between a Picarro dried with Nafion and a LGR for flux measurements and found an R^2 of 0.99 for CO_2. While the correction is not perfect, it seems sufficient for our purposes, particularly that we are attempting to determine the accuracy and precision of the NDIR sensors relative to the LGR, not to the absolute concentration.

Additionally, previously this instrument's water vapor correction was evaluated, but was focused on CH_4 by adding different moisture into the sample line with a calibration gas flowing. The measured [CH_4_dry] is constant at different RH values. This experiment was not designed for CO_2 assessment because the calibration gas went through a bubbler at different flow rates and some CO_2 dissolved into the water.

You calibrate the instrument with two points by assuming a linear fit. How do you know the true instrument response is linear?

A previous calibration for this instrument was done with five NIST-traceable standards ranging from 369.19 to 516.41 ppm which showed its linearity. This has been clarified in the text.

How did you decide to measure the tank at 23 and 47 hour intervals? How do you know that significant drift did not occur over shorter intervals?

We wanted calibrations to occur at different times throughout the day, thus the 23 and 47 hour intervals, and was not performed more often in an attempt to have enough calibration gas for the entire period. As described below, we were unsure how long we would need to equilibrate/get an idea of the repeatability of the LGR instrument.

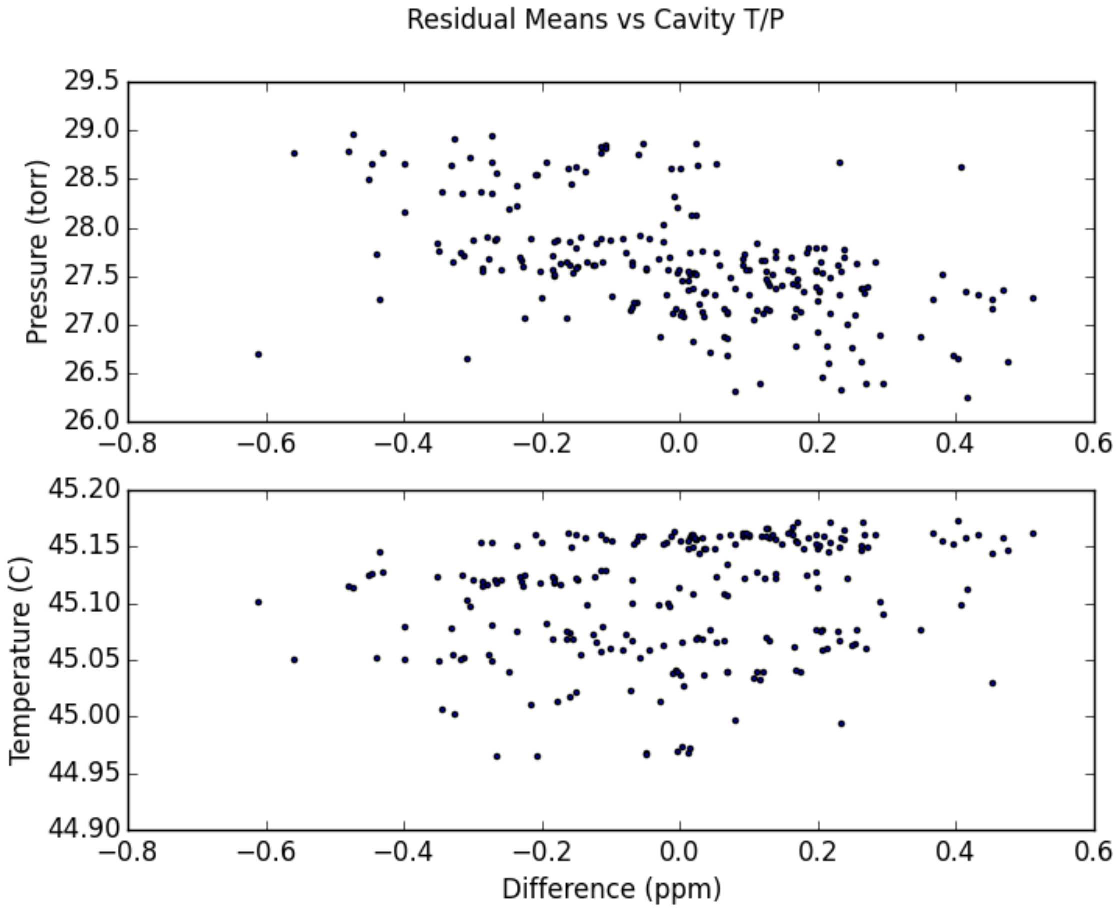
Why do you need to measure the tank for such long time periods (10-60 minutes)? Does it take that long for the measurement to equilibrate? If you are using the proper materials in your plumbing, the measurement should equilibrate in a matter of seconds to minutes. If it is taking a long time for the CO2 signal to equilibrate, that suggests that CO2 may be absorbing/desorbing onto the walls in your plumbing.

The tank is located downstairs indoors about 5 meters or so from the LGR, so there is some time required for the tubing as well as the LGR cavity to flush. Looking at the raw data, it takes somewhere on the order of 90-120 seconds to fully equilibrate. We were initially unsure how long the LGR needed to equilibrate, and wanted an idea of the variability/stability at the fixed concentration, which is why we initially ran the calibration gas for 60 minutes. It was switched later to conserve the breathing air tank.

What is the purpose of the Dasibi calibrator? Did you have to dilute the tank air to get ambient values?

The Dasibi calibrator is purely used as the scheduler for the tank used in the LGR stability evaluation. It contains a clock that turned the calibration gas on/off at the specified intervals. No dilution was performed, as the tank of breathing air provided a concentration within the normal range observed.

Figure 3: If you take the linear trend out, are the remaining variations related to a physical parameter such as temperature (ambient or cell)?



The mean residuals from the linear trend plotted against the mean cavity temperature and pressure during the calibration periods show some correlation with pressure (R^2 of 0.27) and virtually no correlation with temperature (R^2 of 0.06)

Your drift correction technique of fitting a line to each subsequent pair of calibration points will introduce discontinuities into the corrected dataset that do not represent the real-world. It would be better to fit a smoothed curve (captures short-term drift) or a single linear fit (captures the long term drift).

Thank you for the excellent suggestion, using the single linear fit actually improves the RMSE slightly across all the sensors. The analysis throughout the paper (figures, table, numbers) will be update to reflect the results with the linear correction of the LGR drift.

When describing the differences between the K30 and LGR, you say that the LGR cavity temperature and pressure are relatively well controlled. Please give some numbers to give us a sense of how well controlled they are.

Over the entire evaluation period, the standard deviation of the 2-second data is 0.44 torr for cavity pressure and 0.06 ºC for cavity temperature. This has been added to the manuscript.

You average the datasets into 1-minute bins based partially on the Allan variance results for the K30. Did you also do an Allan variance for the LGR?

Doing an Allan variance on the 0.5Hz LGR data using the breathing tank reveals that the noise is also Gaussian and that the optimum averaging interval is ~100 seconds, so 1 minute is also appropriate for the LGR data.

K30 Sensor Performance and Evaluation

Sect 2.1 – Did you compute an Allan variance for more than one sensor? Do they all perform similarly?

Yes, they all perform similarly. This has been added to the paper.

Figure 4 – CO2 traces show periods of higher noise on some sensors (e.g. K30- 3 during the second half of the time period shown). In particular, I am wondering about the smattering of points that appear as outliers. In these cases, are the sensors still meeting the manufacturer's specification of +/- 30 ppm? Is there evidence that a sensor's precision can diminish over time?

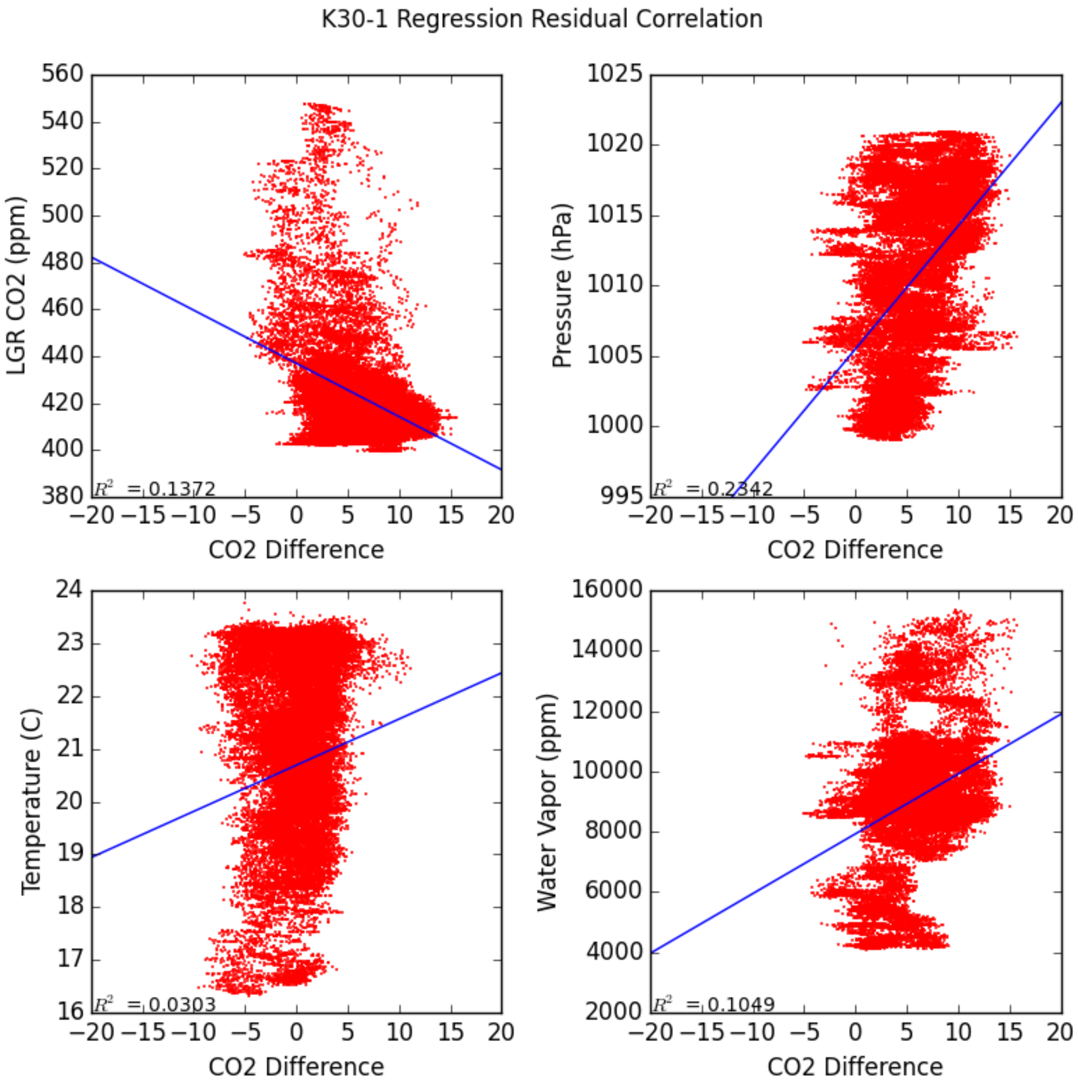
Yes, there are some outliers of the two poor performing sensors that are outside of the ±30 ppm range from the original dataset, but after accounting for the zero offset, only K30-3 has any values outside of this range. This particular sensor has much more noise compared to the others, but for 1-sigma, it is within the specifications.

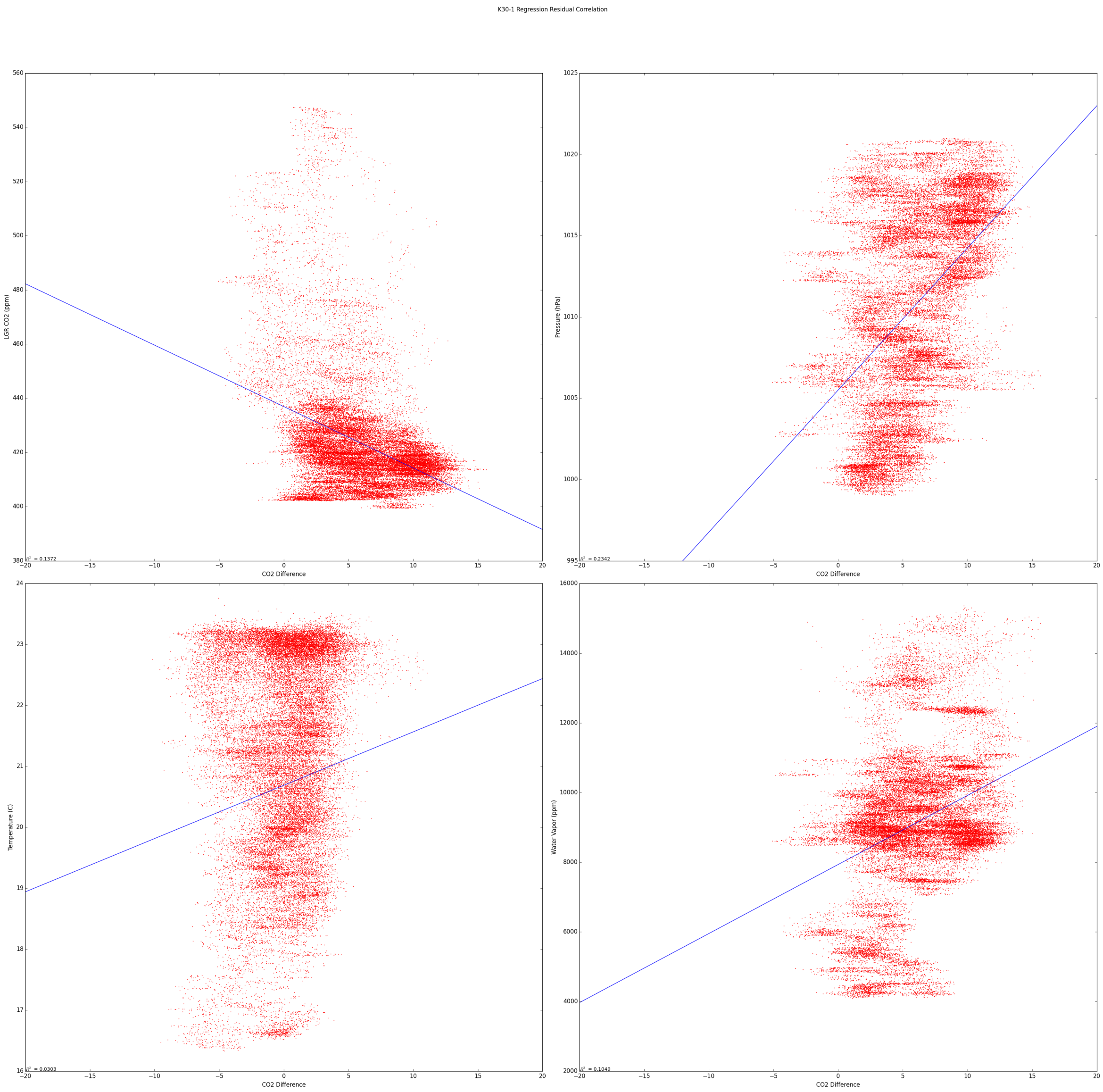
It's too difficult to say from this dataset if there is any evidence of precision diminishing over time. There is definitely a possibility as dust that could collect on the internal mirrors could change the absorption/path length of the IR light, and the light source could also potentially change with use. This would be something we hope to investigate in future work by performing a long-term (6-12 month) evaluation.

At the beginning of section 5, you state that CO2 measurement differences are correlated with environmental variables, but you have not demonstrated the correlation. Can you show some scatter plots?

The correlations are not perfect by any means, but shown below are scatter plots for K30-1 for before the four stages of the regression:

For each plot the x-axis is the original K30 data – LGR. Top left: the difference versus the LGR values, top right: the difference versus atmospheric pressure, bottom left: difference versus temperature, and bottom right: difference versus water vapor. Temperature is the least correlated (the K30 includes a crude first order temperature correction) and is the variable that has little effect on the RMSE (at least in this range of observed values), and pressure is the highest. Note the density of points are not well resolved in the figures and will skew the fit lines, this is shown below with the second plot (larger figure).





Before doing an empirical fit to the environmental parameters, it would seem sensible to account for the dilution of the CO2 mixing ratio in humid air. See section 2 of Shusterman et al. 2016 for an example.

The multivariate regression takes into account the water vapor mixing ratio, as well as temperature and atmospheric pressure, so this should be accounted for when regressing against the CO_2_dry output from the LGR. The correction described in Sect. 2 of Shusterman et al. 2016 is essentially a simplified version of the multivariate regression where they correct for varying T,P,q.

Table 1 – Are all of the regression coefficients significant? Which parameter leads to the biggest improvement and which leads to the smallest improvement?

The most significant correction comes from the simple regression against the LGR reported CO2 values. Otherwise, because the sensor uses the absorption of infrared light, from the ideal gas law, it relates the concentration relative to a reference pressure, so atmospheric pressure has the largest correction. If you change the order, the final result is still the same, but since T/P/q are all correlated from weather and diurnal variations, the first one can often have the most significant impact.

Section 6.1 – Do you find that the K30 sensors that were closer to the LGR inlet have shorter lag times relative to the LGR response? You should try computing cross- correlation functions for each K30 against the LGR to improve the time-matching of the different time series.

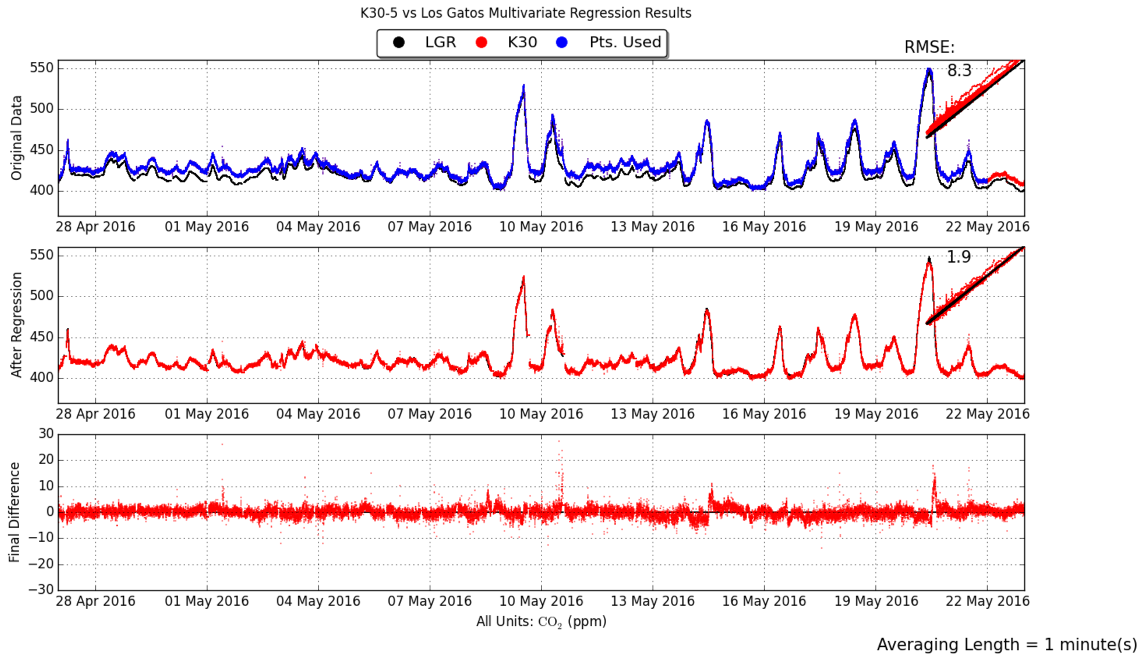
Unfortunately, we do not have an exact record of the location or distance from each K30 to the LGR inlet. The cross-correlation functions would be difficult on this dataset because the lag can vary depending on things like time of day and weather, thus the lag correction may change throughout the time series. In a real-world setting, there could be occasional dramatic shifts in concentrations from plumes, boundary layer dynamics, or other reasons, so this demonstrates that there is a need to wait several minutes until the signal equilibrates somewhat.

Significant figures - Most of the performance metrics stated for the K30 sensors in units of ppm CO2 are given with two decimal places, yet you state the K30 measurement resolution is only 1 ppm.

Yes, the output is 1 ppm resolution, but for 2-second data, when using the 1-minute averages, the effective resolution is higher. We have now changed these units to use only one decimal place, which is more consistent with the precision of the 1-minute averages but still shows the impact the regressions have on the RMSE.

Figures 5,6,7,8,10 are all shown for K30 #1, which, from Figure 4, appears to be one of the best performing sensors. I would be curious to see a residual plot for sensor #3, 5 or 6.

Here is Figure 8, but rather for K30 #5 than K30 #1.



You state that one goal of this work is to understand how correction coefficients can be derived quickly. Wouldn't it be more efficient to design a controlled experiment where controlling variables are deliberately varied across the full range of operating conditions?

We were looking for a way to derive the coefficients for a large group of sensors with minimum human labor. By artificially controlling air temperature, pressure, and moisture content, the cost of the evaluation both in time and money would increase, negating some of the benefits of the price of these sensors.

Section 7 – In future work, you aspire to characterize the sensors' maximum performance in a controlled environment. Yet, if the big-picture goal is to use these sensors is to generate science quality ambient air measurements, I believe a more worthy goal would be to characterize their minimum performance in an uncontrolled environment.

This is another area we hope to pursue with future work, but would need to devise a way that would be uncontrolled but also meaningful enough for publishable data. Ideally we would like some installed outdoors next to an inlet for a gas analyzer as the reference, but would need to ensure that the sensors are in an enclosure that provides adequate ventilation but also protection from weather.

Technical comments

Title – I don't think "enhancement" is the right word. How about something like "Evaluation and correction of CO2 measurement in ambient air from a low-cost sensor"

Thanks for the suggestion, we changed it to "Evaluation and environmental correction of ambient CO_2 measurements from a low-cost NDIR sensor"

Abstract – The quantities reported have different numbers of significant figures. These should be uniform and reflect the precision of the measurement.

As mentioned above, we now use one decimal place for the RMSE of the K30 sensors.

Pg 1, Ln 25 – "dry air" is used twice in this sentence.

Fixed.

Pg 1, Ln 28 – The WMO compatibility goal is a goal, but is not always achieved, and certainly not for historical measurements.

We agree with the reviewer. We have changed this to state that this is the WMO compatibility goal.

Pg 1, Ln 29 – Suggest: '. . .to collect samples, which are subsequently transported'

Thanks, changed it to reflect this suggestion.

Pg 1, Ln 30 – You mention two expensive types of measurements – flasks and Picarros, but you do not mention moderately-priced analyzers from LiCor and Los Gatos, which are used at many research-grade monitoring sites.

This is true, but the added costs for calibration and maintenance can make a LiCor (and Los Gatos as you've seen from our analysis) comparable in total cost to a Picarro. The models are not explicitly stated and this is merely to show that either flasks or continuous observations are prohibitively expensive to do at high spatial resolution. We have added a phrase to indicate that this cost includes labor and calibration costs.

Pg 2, Ln 1 – Be consistent about whether you spell out carbon dioxide or use the abbreviation.

This has been corrected throughout the manuscript. Thanks.

Pg 2, Ln 2 – Suggest paragraph break at "Recent research"

Done.

Pg 2, Ln 8 – Is 8-12 sites typical? There are ~5 sites in Boston, SLC, and Paris.

For the LA Megacities project there are 14 sites, 11 current sites in Indianapolis, and a planned 14 sites for DC/Baltimore, 8-12 was used as a rough average of these 6 cities. The text has been changed to 3-12 to reflect the inclusion of the smaller networks listed. Additional references have been added here to show a sample of these networks both in size and geographic location.

Pg 2, Ln 9 – You say that more dense observations, even with larger uncertainties, yield better inversion constraints, but this is all relative and depends on the inversion setup/goals. See Turner et al., 2016, ACP for an exploration of the tradeoffs.

The text has been updated to state that this depends on the methodology used, and a citation to Turner et al., 2016 has been added.

Pg 2, Ln 14 – Suggest deleting "however"

Done.

Pg 2, Ln 16 – Suggest changing the phrasing to: "Recent evaluations and implementations of new low-cost sensors demonstrate their promise for ambient air monitoring."

Changed to "Evaluation and implementation of some of these new low-cost sensors demonstrate their promise for ambient air monitoring."

Pg 2, Ln 26 – Can you give some numbers to scope what you mean by "reasonably accurate"?

Based on the cited texts, ±3-5ppm, has been added to the manuscript.

Pg 3, Ln 7 – Suggest: "The K30 sensor module from SenseAir (Sweden) is the low-cost NDIR CO2 sensor that was tested for this study".

Changed to "The K30 sensor module (K30) from SenseAir (Sweden), is the low-cost NDIR CO_2 observing instrument used in this study[1]."

Pg 3, Ln 10 – Suggest deleting "given as"

Thanks. Changed.

Pg 3, Ln 13 – Suggest: "The K30 was chosen not only because it has the highest manufacturer-specified accuracy, but also because initial testing showed reliability and consistency with higher-quality observations."

Changed to close to your suggestion: "The K30 was chosen not only because of it has the highest manufacturer-specified accuracy, but also because initial testing showed reliability and consistency when compared to higher-quality observations."

Pg 3, Ln 17 – You should give the units (relative humidity) for the 3% and 0.008% quantities.

Thanks for the suggestion. We have revised the sentence as "…has an average absolute accuracy of ±1 ºC, ±3 %, and ±1 hPa, and an output resolution of 0.1 ºC, 0.008 % and 0.01 hPa for temperature, relative humidity, and pressure, respectively"

Pg 3, Ln 24 – "less than one percent" "< 1%"

Fixed.

Pg 4, Ln 5 – Another difference between the two analyzers could be their sensitivity to the isotopes of CO2.

This could be true, but without knowing for sure, we prefer to not add this point to the paper. The LGR is only sensitive to ^{12}C, but the standards used to calibrate the LGR account for all isotopes of CO_2. Additionally, the component of ^{13}C is around 1% relative to ^{12}C, and thus the difference would be small. This is now briefly addressed in the text.

Pg 4, Ln 29 – Can you briefly describe what you mean by "various complications"?

The Raspberry Pi runs a full Linux OS, so because of the complexity of the OS, sometimes there is a delay in when certain tasks execute, which may compound into some sensors collecting (at times) observations out of phase from others. The LGR 0.5Hz data starts whenever the system initializes. Thus perfect synchronization is difficult, but all have recorded time stamps and can be averaged / regularized for comparison. A brief explanation of this has been added to Section 2.

Pg 5, Ln 1 – My understanding is that you merged datasets by their timestamps. Did you have to do something to keep the clocks synchronized?

All of the Raspberry Pi data loggers use an internet server to synchronize their time, and the LGR uses an internal clock with battery that was set to the same time as the Pis at the beginning of the experiment. This has been added to Section 2.

Pg 5, Ln 13 – "longer averaging times do not reduce the noise"

Fixed. Thank you for noticing this.

Pg 7, Ln 21 – Suggest deleting "However".

Done.

Pg 7, Ln 22 – Is the statement about each K30 meeting the manufacturer's uncertainty specification in regards to the raw (2-second) data or 1-minute averages? Please clarify in the text.

The manufacturer specifies the range for the raw data but our analysis is for the 1-minute averages. Text is updated to reflect this, both in section 2 that the datasheet is for 2-sec and in section 4 that our analysis is for 1-min.

Pg 10, Ln 21 – suggest: ". . . and 1.48 ppm, for 1-minute, 10-minute, and hourly averages, respectively.

Text is changed to this suggestion.

Pg 10, Ln 26 – suggest: "One goal of this work is to develop a methodology to evaluate individual sensors quickly. . ."

Changed to "One goal of this work is to develop a methodology to evaluate individual sensors quickly so that they can be used in scientific applications."

Pg 11, Ln 32 – "less than five parts per million" "< 5 ppm"

Fixed here as well.

Figure 1 – A ballpoint pen is included in the picture for size reference. A ruler instead of a pen would be more useful.

We liked this idea, and Figure 1 now includes a ruler instead of a ballpoint pen.

Figure 2 – What was the CO2 concentration of the tank used?

This is the breathing air tank used for the LGR drift, so estimated to be 463.7 ppm after calibrating the LGR with NIST standards, as noted in the text at the end of Section 3.

Figure 4 – State the time interval of the data shown. I can't tell if this is raw 2-second data or 1-minute averages.

1-minute averages, all figure captions have been updated to clarify this.

Figure 8 – I don't understand the difference between the red and blue points.

The blue data points are used in the regression, and the red is the complete dataset. This is done for consistency with internal plots that show the time series for regression periods of varying length. Captions for Figs. 8 and 10 have been updated to clarify this.

Figure 9 – Can you put error bars on each point for the y-variable?

We decided to instead show a box plot as well as all six sensors' values, as this gives a better picture of the variability than just error bars. Please see updated Figure 9.

Response specific citations:

Yang, M., Prytherch, J., Kozlova, E., Yelland, M. J., Parenkat Mony, D., and Bell, T. G.: Comparison of two closed-path cavity-based spectrometers for measuring air–water CO_2 and CH_4 fluxes by eddy covariance, Atmos. Meas. Tech., 9, 5509-5522, doi:10.5194/amt-9-5509-2016, 2016.