Fleming and coauthors present an evaluation of a commercial cavity ring-down spectrometer, both in terms of its ability to meet WMO compatibility criteria for O2/N2 under specific conditions, and its suitability for in situ measurements.

The suggested advantages of the Picarro analyzer are that it can be run without sample drying and does not require continuous reference gas flow. This would make it attractive for installation at a remote site. However, the authors show that the instrument is unsuitable for such an application, due to the large artifacts the analyzer is subject to under such conditions. As the authors show, it does perform reasonably well when measuring tanks. For this reason I think the authors focus a bit too much on the compatibility/repeatability goals of the Picarro under laboratory conditions. Much more telling is the in situ data. I am not sure the fCO2 discussion adds much to the paper, since it relies on CO2 measurements not made by the Picarro analyzer. The in situ comparison presented is more to the point, and sufficient for the demonstration. The authors could even cut the fCO2 comparison from the paper to reduce the length, in my opinion.

I think this is an excellent paper and of interest to the AMT readership. I recommend publication with only minor comments.

Minor Comments

L17 and throughout: It would be easier on the reader to stick with a single unit, rather than switching between ppm and per meg.
L17: I think the wording needs refining here, do you mean that the highest precision possible was found at 300 seconds? What does it mean to report an Allan deviation as 1 standard deviation? For an abstract I think it's sufficient to say that you estimated the precision to be 1 ppm, reported as 1 sigma, from 300 second means.

L21: pre-dried is confusing, suggest "dried". The grammar is a little off in this sentence due to the mixing of tenses.

L24: The abstract is quite long, suggest cutting "(sometimes known as a "surveillance tank")"

L43: Better to give the increase of CO2 over the same period as O2 (past three decades).

L55: I think it's confusing to give an approximate definition for APO when saying it is defined as, better to give the actual equation.

L62: It would be good to define compatibility here...in L66 it seems conflated with precision. Doesn't compatibility combine accuracy and precision into a single metric?

L81-84: To measure O2 with high precision and accuracy you need all of these things. The author is suggesting that they can all be contained within a single box, which is certainly convenient. "Revolutionize" seems a bit too strong to me. There might be some savings in avoiding the continuous use of a reference gas, but the Picarro analyzer is expensive, and all of the other expensive, labor intensive aspects to making in situ measurements would still be needed.

L89-91: And yet the authors go on to show that the instrument does NOT have all of these advantages. I think this needs some rephrasing..."the vendor suggests that" or "it is intended for" use without drying, cal gas, etc. To be fair to Picarro, maybe this is not what they had in mind. There are other applications for this instrument beyond the small field of high-precision atmospheric monitoring.

L125: What's the flow rate, and how big is the cell? Does it really take 8 minutes to flush it? This is an extremely long e-folding time. It would be nice to see some of the actual calibration data. If the sample air is wet and the calibration gases are dry, isn't it more likely it's a surface effect rather than a purging issue?

Figure 1: How is pressure/flow control maintained for the Oxzilla? I see no pump depicted.
L173: change "scales," to "scales. This"

L177: There are also surface effects to consider, the dilution effect is not the sole reason.

L202: Really? Again, I find this surprising.

L250: Please give +/- on cavity pressure, temperature, and flow.

Figure 2: It would be nice to see the x-axis extended here, since the time horizon for the RT (5 hours) is outside of the plot.

L366: I don't fully understand this. It looks like the grey points are jumping at calibration intervals, and the calibration coefficients were not interpolated between calibrations, but applied stepwise. Why would the Picarro instrument's baseline jump always at calibration times? This is actually the only shortcoming of the paper--it would be nice if the author's could speculate more as to what is going on to cause these baseline shifts.

Figure 8: I think it would be better to drop the no RT Picarro data here, zoom in on the y-axis, and make the points open circles (and smaller)--it is hard to see the data which matters, which is the Oxzilla vs the RT-corrected Picarro.

L400: "which provides a measure of the compatibility to the SIO O2 scale over time" -- I'm not sure that's quite correct, unless the tanks are being remeasured at SIO for each comparison.

L455: Maybe it's worth pointing out here (or earlier) that for in situ measurements, the O2/N2 ratio will be changing over tens of minutes. Averaging down pure random noise is not the same as averaging observations over an hour.