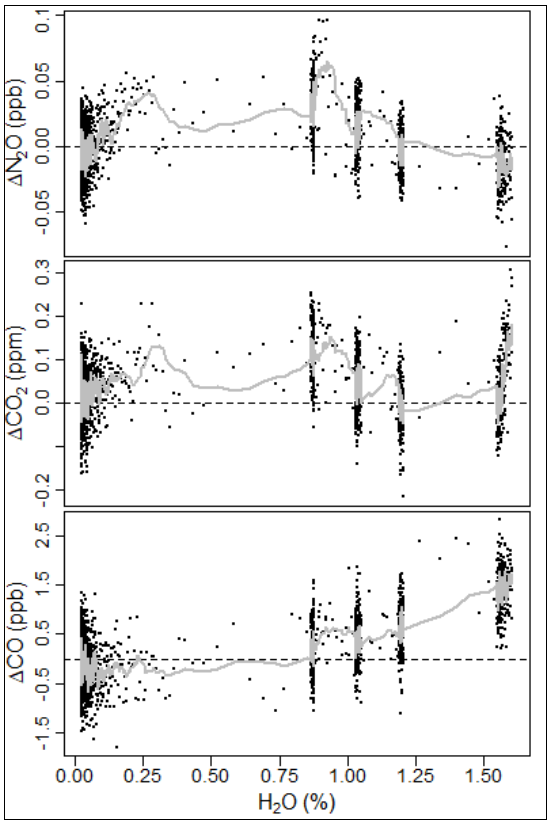
Thank you for the thorough review and suggestions. We have responded to each comment in-line below and updated the text—adding more text on key issues raised and including more figures to support the discussion. We believe the manuscript is now much improved.

J. R. Pitt:
It has now become well-known in the trace gas measurement community that analysers employing Tunable Infrared Laser Direct Absorption Spectroscopy (TILDAS) techniques on aircraft can exhibit a strong sensitivity to changes in cabin pressure. This is currently a major limitation to the utility of these instruments for airborne sampling, as operators either have to accept large altitude-dependent biases in their final dataset or very low duty cycles (as the instrument must be recalibrated at each altitude). Many important trace gas species (e.g. N2O and C2H6) have much stronger absorption lines in the mid-infrared than the near-IR; whilst cavity-based measurement techniques for these species have developed significantly in recent years, TILDAS is still the most commonly used technique for measuring many trace gases in the mid-IR. Therefore improving the accuracy and duty cycle of TILDAS instruments during aircraft sampling is important if we are to improve our understanding of key greenhouse and pollutant gases.

This paper presents a novel calibration strategy to tackle this issue, resulting in greatly reduced altitude-dependent biases with a 90% duty cycle. The switch to controlling mass flow instead of pressure is a clever idea, removing issues associated with pressure instability during the sample-calibration transition. The method employed here will certainly be of great interest to anyone currently operating TILDAS analysers on aircraft, but it also provides the potential for reducing biases and/or increasing the duty cycle for other airborne instrumentation. I recommend its publication in AMT, but I have the following suggestions for minor revisions.

The water broadening correction is mentioned at the end of section 2.1. Did you determine the water broadening to air broadening ratio yourself experimentally? If so perhaps state this explicitly because at the moment it makes it seem like TDLWintel does this automatically without user intervention (unless Aerodyne tested your instrument before sending it to you – in which case mention this because as far as I'm aware it isn't what they don't usually do). Assuming you determined the coefficients yourself, could you add a brief outline of the general approach taken (e.g. H2O injection/dew point generator/etc. . .) and the uncertainties associated with it please? The uncertainty associated with this water vapour correction is often on the same order as the other uncertainties so it's important to consider it.

Thank you for raising this point – this is of great importance to accurate trace gas observations and we have added a section discussing this. We now describe the water broadening coefficients, our determination of the values, and the associated uncertainties, and we have added a figure showing the effect of the correction as a function of water vapor. We used a moist filter similar to Lebegue et al. 2016, and sampled a mixture of wet and dry tank air which we then reanalyzed to find the proper coefficients. Table 1 has also been updated to include these uncertainties. We also note that in adding this discussion we realized the proper water broadening coefficients were not consistently applied to the data presented initially. We have corrected this oversight for this dataset in all the data and figures presented in the revision.



Section 2.3 is presented in a rather confusing way. I think this is largely because the use of two in-flight calibration gases is introduced here, but the fact that one of them is used as a check gas is not mentioned until section 4.1. It wasn't until I got to this point much later in the text that I fully understood what was going on (e.g. why the long-term drift in instrument slope needed quantifying in the lab) – it would be much better if it was explicitly stated in section 2.3 that a single-point calibration strategy with an additional check gas was used in-flight. Additionally, the term "linearity" is used in a way here that isn't intuitive to me. I would stick to using the word "slope" throughout (or "gain" if an alternative is needed), as what is being tested here is the extent to which the instrument slope drifts with time. In my mind instrument linearity is the extent to which the linear fit used here is applicable, not whether the coefficients are drifting with time. To assess linearity three cylinders with different mole fractions are therefore needed. This hasn't been done here, but these instruments are known to have a good linear response so it's probably safe to assume non-linearity is a small component of the uncertainty budget.

We have updated Section 2.3 to clarify this point as suggested. We have also changed the use of "linearity" to the more appropriate "slope" throughout.

Section 3.2: having had many discussions with Aerodyne about this following on from our 2014/15 campaigns, I am fairly certain that the main source of the large gradients in mole fraction you see during profiles is indeed an optical fringe (or possibly multiple fringes), activated by the change in cabin pressure. However we do also see artefacts in the measurements

associated with aircraft acceleration as well – for us these manifest themselves as much smaller-timescale features which would not be captured by the calibration strategy here.

Optical fringes are likely the dominant component. Acceleration-induced artifacts can have very short duration (g-force), and not be well-captured by the method here. They can also have a longer duration (e.g. large engine pull for the duration of a climb possibly impacting instrument via electrical feedbacks), and these would be captured. As shown in the paper, longer-duration features dominate our artifact and thus are the focus of improvements.

We have added clarification to the end of first paragraph in section 3.2 "Artifacts that occur on shorter time-frames, such as induced by a short duration turbulence event, will not be corrected with this method."

Section 4: our experience is that there is significant flight-to-flight variability in the cabin pressure artefact, making it impossible to apply a single correction throughout a campaign. This is because, while the FSR of the fringe at ground level stays constant, its initial phase is variable and unpredictable. We did experiment with doing two deep profiles at the beginning and end of a flight on tank air to try and calibrate out this effect, but the drift during the flight resulted in a large uncertainty and we abandoned this approach. Essentially I'm not convinced that this effect is repeatable enough to take data from a single null flight and use it to correct other flights, so I'd probably remove that paragraph (unless you have evidence to the contrary). The method you've developed here seems far superior to any that could be developed from the null test results.

We agree that our method is an improvement on attempting a single cabin pressure artifact correction. We've modified the text in Section 4 as outlined below to clarify.

"Given the repeatable, smooth nature of the cabin pressure artifact, it *would seem* possible to use just the cabin pressure data to empirically correct for the artifact, without running frequent calibrations. This method *would not* account for long-term spectral drift however or traceability, *and relies on the assumption that the cabin pressure artifact will be stable and repeatable. These weaknesses compromise such an approach.*"

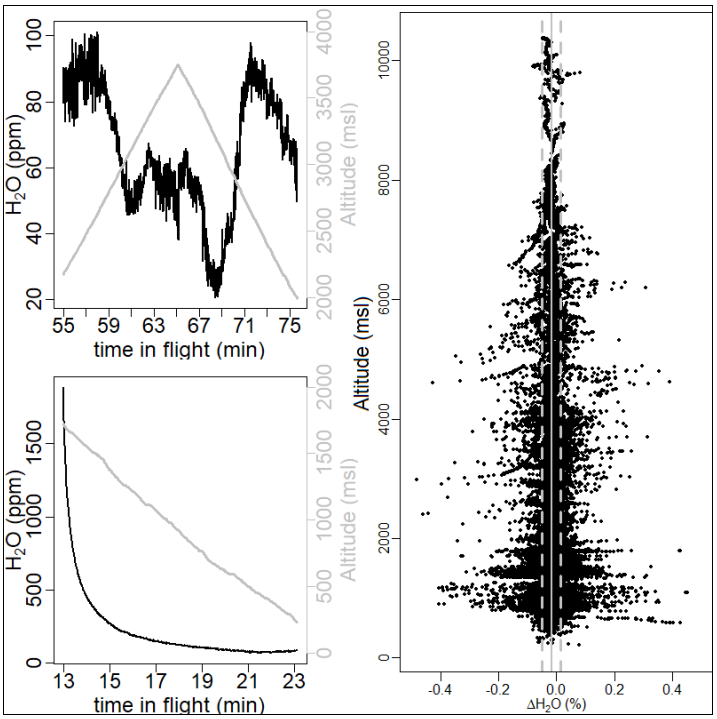
I'd be interested to know more about the suspected contamination which has resulted in the 0.6 ppm CO2 offset between the Picarro and the FCHAOS systems. Have you been able to identify which of the systems the contamination is associated with? Reading the manuscript I initially assumed that it was the FCHAOS regulator/tubing that was under suspicion, but it would be good to make this explicit, or if it is still unknown which instrument was contaminated to state that. Was there any change to the setup in future campaigns where this offset was not observed? The regulator and tubing used for the FCHAOS here are pretty standard, so if there are contamination effects associated with either it would be good to know about them! Or is the theory that the cylinder itself became contaminated (e.g. due to a mistake during regulator flushing)? Surely in that case you would also expect to see the offset in the check cylinder data in Figure 8? In the first half of the campaign (before the cylinder switch) there may be a sign of a negative bias in CO2, but there doesn't appear to be a corresponding positive bias in the second half, so this doesn't really tally with a single cylinder contamination. Also in that case you'd

expect the Picarro bias only to be present in one half of the campaign but I can't see any evidence of this in Figure 7. If you haven't made any further progress in diagnosing this then no worries, but it would be good to include any extra details you do have.

We appreciate your interest in this cause. We have gone back through in further detail trying to isolate a potential cause, but still are left with a somewhat unsatisfactory suspicion of contamination. We have modified the text in Section 4 slightly in accordance with this. We also add the discussion here to further highlight the importance of the water vapor correction you raised earlier, noting that residual water vapor errors (in either the FCHAOS or Picarro systems) could contribution to perceived biases.

The fact you don't see the same effect on the H2O measurements is very interesting, but I can't quite see this from the plots included here. Could you add another column in Fig. 4 showing H2O please? I know the tanks were dry but you are assuming (in my view reasonably) that the artefact is a simple offset shift so it shouldn't matter what the absolute value of the H2O mole fraction is – even if it is completely dry I'd expect the fringing to affect this zero-offset reading. I don't doubt your word here, but the explanations offered as to why the fringe would affect H2O less don't really make sense to me either so I could do with a bit more detail on these. Surely the fringe amplitude will increase with laser intensity, if anything making the problem worse? The relevant signal-to-noise here is the strength of the absorption peak (not the laser intensity) relative to the fringe amplitude. If the H2O line in question is the one at ~2227.5 cm-1 then this tends to be a weak feature relative to the N2O and CO2 lines, so again I would have thought that H2O would be more affected by the fringe. I'm also not sure why having a line frequency of ~2227.5 cm-1 compared to the CO2 line at ~2227.6 cm-1 (for instance) would reduce the fringe interference. It is definitely true that a very wide absorption feature would suffer less from a fringe with a small FSR, but I don't think the H2O line is wider than the other peaks here? Sorry to labour the point on this, but the fact that you don't observe the fringe effect on the H2O measurement could be a really useful piece of information in trying to further mitigate this issue, so I'm keen to better establish the cause of it.

Thank you for digging into this question further. We have updated Section 4 when discussing comparisons between FCHAOS and the Picarro, and added the figure below to show the water vapor during the null test profiles as well as difference between FCHAOS and Picarro measured water vapor as a function of altitude. The long equilibration time for water vapor makes it difficult for us to see an altitude artifact if it is present. One null test profile suggests there may be a ~60 ppm water vapor vertical sensitivity, but it is difficult to clearly establish. In the comparison with the Picarro we see no evidence of any vertical dependency of water vapor, though a dependency of 10s of ppm would be hidden within the respective instruments noise. We have updated the text to include this more nuanced discussion.



Section 5: Could you put details of the altitude and variability in altitude during the runs in here please? I assume it was essentially performed at a single level but if so it would be good to be clear about this just so the reader knows there are no vertical gradients convolved in here.

We have updated the text with the altitude values, affirming that the transects were at relatively constant altitude.

Specific points:

P3 L2 – Minor point but the LGR FGGA in O'Shea et al. is a near-IR instrument

Thank you for catching that, we have updated the text for accuracy.

P3 L26 – Typo: missing space

Corrected.

P5 L18 – ". . .within our 1 Hz precision. . ."

Corrected.

P7 L9 – ". . .at the same flow-rate. . ."

Corrected.

P7 L12 – What interpolation technique was used?

Forsythe, Malcolm, and Moler (FMM) cubic spline interpolation, we have added this to the text.

P8 L8 – ". . .within our 1 Hz precision."

Corrected.