Author's response to review 2

The paper describes a unique and valuable dataset collected in a harsh environment in an under-sampled region. These data will be valuable for inverse modeling to estimate emissions and removals of CO2 and CH4. Arctic data such as these are particularly needed, since release of carbon from permafrost is an expected outcome from warming temperatures and current estimates of Arctic fluxes vary widely. The authors provide a useful and complete description of challenges of operating in the Arctic and their strategies for maintaining continuous operations and filtering data to remove local effects. The description of the configuration is comprehensive and clear. The authors have provided quantitative and time-varying uncertainty estimates and a clear description of how the uncertainty was estimated.

A concern is that the data is available "on request" rather than readily available for download (e.g. from the WMO Global Atmosphere Watch World Data Center for Greenhouse Gases or these data could be included in the GLOBALVIEW+ ObsPack product compiled by NOAA). The value of these data will only be realized when combined with other datasets from the global community.

We agree with the reviewer that the value of our dataset for the atmospheric research community will be substantially increased by making the data 'visible' in one of the commonly used online repositories. A publication of these datasets in a public and visible repository (e.g. WDCGG) is therefore foreseen for the near future.

Also, the spike detection algorithm seems to be highly tuned and somewhat arbitrary (but to be fair data from many sites are manually flagged, which relies on expert judgment that is arguably even more arbitrary). Please see specific comments about making the flagging criteria explicitly available so that users have enough information to develop their own filtering scheme.

All criteria and thresholds used for the spike detection were given in detail in Appendix D, which should enable reproducing the procedure. Since the chosen settings were customized for Ambarchik, we agree with the reviewer that an adaptation of this method at other sites would require an adaptation of these criteria. Still, we believe that we presented an objective method to remove spikes that both clearly demonstrates how we filtered our own data, and moreover should be applicable also to other datasets, given that the PIs are willing to fine-tune the settings.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Review Criteria for AMT: ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Does the paper address relevant scientific questions within the scope of AMT? yes

Does the paper present novel concepts, ideas, tools, or data?yes the data from this new Arctic site are novel and uniquely valuable for tracking possible release of CO2 or CH4 from permafrost

Are substantial conclusions reached?yes in the sense that 2+ years of data are presented along with an assessment of enhancements over background presented versus wind direction and season

Are the scientific methods and assumptions valid and clearly outlined?yes
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Are the results sufficient to support the interpretations and conclusions?yes

Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? yes with some minor requests for clarification below

Do the authors give proper credit to related work and clearly indicate their own new/original contribution?yes

Does the title clearly reflect the contents of the paper?yes

Does the abstract provide a concise and complete summary?yes

Is the overall presentation well structured and clear?yes

Is the language fluent and precise?yes

Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?yes

Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?no, the paper is of appropriate length and detail

Are the number and quality of references appropriate?yes

Is the amount and quality of supplementary material appropriate?yes

We thank the reviewer for this very positive evaluation of our manuscript in light of the AMT review criteria.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Specific Comments: ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

page 6 what is the flow rate through the analyzer and what is the purge flow rate?

The nominal flow rate in the sample line (as measured by FM2) is ~170 mL/min. We add this information to Sect. 2.3. The purge flow (FM1) is ~17 L/min and was already reported in Sect. 2.3.

page 9, line 25: Is there any indication if the time synching with the GPS fails?

Time synchronization takes place between GPS clock and Picarro clock, and separately between Picarro clock and data logger clock. The latter synchronizations are protocolled and can thus be checked. However, this is usually not done, since so far there was no indication that there were timing issues.

page 10 line 20: State that "synthesis" function is defined in Appendix B.

We add this reference there.
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

page 10 lin3 16: Variability of water correction experiments discussed by Stavert et al., AMTD, 2018 (https://www.atmos-meas-tech-discuss.net/amt-2018-140/) and could be referenced here. They found that short-term repeatability of water corrections was similar to long-term repeatability.

This could indeed be an indication that we observe short-term variations. We add this information to the text.

page 12: what is the expected lifetime of each calibration cylinder?

The Target tank was replaced in summer 2018 and thus lasted 4 years. At this rate, regular calibrations would deplete the calibration cylinders in 16–24 years. However, these tanks are also used for other experiments like water corrections and are thus depleted far more rapidly. They are currently 40-50% depleted (status: February 2019, i.e. after 4.5 years of operation).

page 13: it would be useful to describe the stochastic and non-random components of the estimated measurement uncertainty (i.e. to what extent does the uncertainty improve with averaging).

This topic is mentioned again in another comment below, where we address it.

The text states that the uncertainty is dominated by the water correction, which is not going to improve with averaging. But perhaps also include a statement about the short-term precision of the analyzer for each gas (i.e. what is the standard deviation on each 10-minute calibration after the gas has equilibrated).

We discussed short-term precisions in the response to review 1, and discuss random and systematic uncertainty components in the response to the review comment about page 32 below.

What is the typical standard error of the residuals?

The standard errors of individual calibration events were 0.018 ppm CO2 and 0.04 ppb CH4.

page 16: description of data filtering algorithms is useful and the results shown in Table 3 demonstrate that impact is practically negligible.

Ok.

page 16: description of water vapor spikes is interesting, and the explanation seems plausible

Ok.

page 17: it would be useful to see how the virtual potential temperature threshold corresponds to other indicators of difficult-to-model observations. For example, are hourly standard deviations typically higher than during well-mixed conditions? What is the duration of a typical inversion (i.e. how many consecutive hours of data are typically flagged)? Can these events be reliably screened based on something like enhancement above a smoothed background? This type of information could be helpful for developing filters for other sites (particularly Arctic sites) where virtual potential temperature information is lacking.

We agree that our data might be useful for developing filters for other sites. However, it would not be guaranteed that relationships between filter criteria in Ambarchik would be valid universally. Transferable relationships would have to be validated with analogous data from other sites. This is beyond the scope of this paper, but could be done with the data we distribute. This may be an interesting topic for a follow-up paper. For this paper, however, such an extended analysis is beyond the scope we set up, so we decided to not follow up on these remarks.

page 18, line 5: what is the duration of the back trajectories (i.e how many hours or days backward in time)?

We used 15-day backtrajectories for these analyses. We add this information to the text.

page 20, line 6: How are Barrow data selected for this comparison. State clearly that you are including Barrow data that has not received a first column flag if that is the case.

Correct, that was essentially the quality filter. We add the following explanatory paragraph to the text:

Barrow data were filtered according to their quality flag. For CO_2, data with quality flags "...", ".D.", ".V." and ".S." were included. For CH_4, data with quality flags "..." and ".C." were included. Data with other flags than a "." in the first column were removed as invalid. Other quality flags (differing in the second or third column) were excluded because their number was negligible.

Can you speculate about why the virtual potential temperature filter would remove such a large fraction of the data at Barrow? Is there some obvious difference in the meteorological conditions at the two sites? Does this result have implications for interpreting the Barrow data?

First, we calculated the intercomparison between Ambarchik and Barrow also without the temperature filter for Ambarchik data, and resulting plots look virtually identical. Accordingly, this filter is not essential for the site intercomparison presented in the manuscript.

As we see it, the differences in the wintertime near-surface temperature profiles between both sites can most likely be related to the surface structure in the near field of the stations, rather than to differences in climate. With Ambarchik being situated close to the shoreline of the Arctic Ocean, on top of a low cliff, the level of mechanically generated turbulence is comparatively high. The Barrow station, on the other hand, is situated in very flat terrain, so that mechanically generated turbulence is less likely to break up stable stratification of near-surface air masses. Thus, temperature inversions might occur less frequently in Ambarchik than in Barrow. Since these considerations are speculative, we do not include them in the manuscript.

page 15: regarding amplitude estimation, maybe it would be better to use the curve including residuals and then estimate the amplitude based on the difference between the min max smooth curve values (and you could just compute the average for all the consecutive min-max or max-min pairs). Then you could do same with to ensure apples to apples comparison. Otherwise when you compare Barrow and Ambarchik are the amplitudes different because of different time periods?

The time periods for estimating the amplitudes for Ambarchik and Barrow were identical, as depicted in the plots. We agree with the reviewer that the min/max fitting outlined by him/her may provide additional information that may become helpful when applying this procedure in other contexts. However, in the case of the data presented herein, we do not believe that this alternative way of computing the amplitudes would add information that would change our data interpretation. We will therefore not change the approach presented here.

page 20, line 22: are you sure that the smaller variability at Barrow was real and not due to differences in screening for the two sites?

Yes, since we used virtually all data at Barrow (see above).

page 22: Were the trajectory endpoints the actual endpoints for the entire Arctic WRF domain? Or did you define a subdomain? It would be useful to provide some information about the locations of the endpoints (such as vertical and lat lon distributions by season or for some typical examples).

We used the domain introduced in Sect. 4 (3200 km x 3200 km). Most trajectories leave this domain, but in case a trajectory did not, its endpoint was sampled within the domain. In Fig. 1, we show the distribution of trajectory end points and their height above ground level for all simulated trajectories. However, we think that this is too much detail for the paper and therefore will not include these figures there.

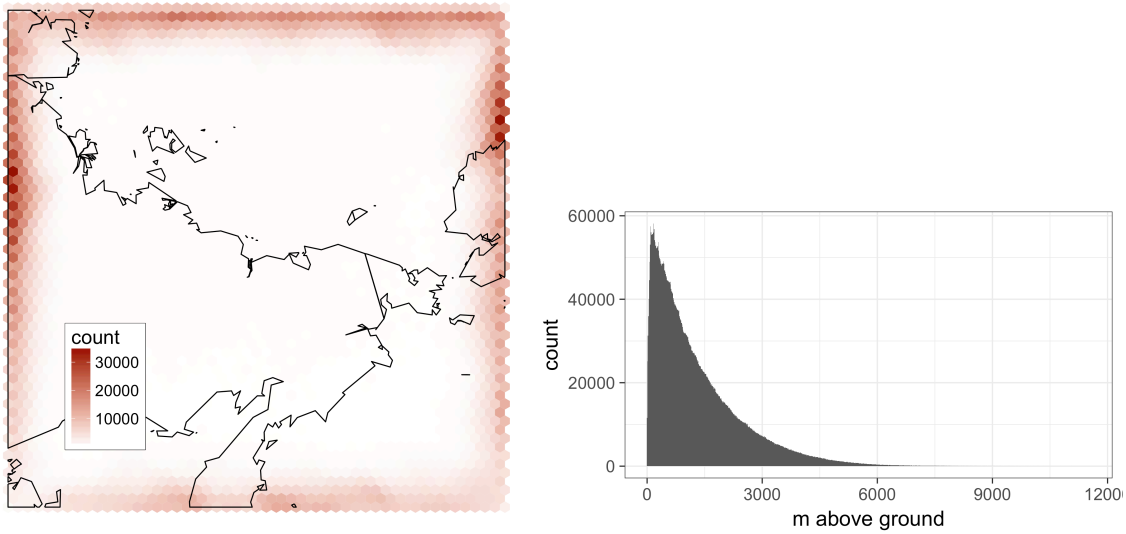

Fig. 1: Spatial distribution of trajectory endpoints for Ambarchik in the domain used in this paper. Left: geographical distribution. Right: height above ground level.

page 24 line 16: instead of "exceeded the goal" perhaps say "did not meet the goal" (although I am not sure the uncertainty estimate is accurate to 0.01 ppm, so maybe you could say instead something like "meeting the goal to within our ability to estimate the uncertainty). Certainly you are doing as well as any other group in the world, and better than most at documenting the uncertainties.

We delete this sentence because, as pointed out by reviewer 1, the values are not directly

comparable.

page 29, line 6: differences among sequential individual co2 measurements?

Yes. We add ‘consecutive’ to this sentence, so that it reads: “Candidates for CO2 spikes are identified based on the variability of differences between consecutive CO2 measurements.”

page 29, line 11: it’s not clear how "cases when all CO2 data in the interval have rather uniform variations" are identified so that they can be unflagged

We clarify by replacing this sentence with the following paragraph:

In some cases, this procedure flags the complete interval as spikes. This happens when the variations throughout the interval are rather uniform. This might be the case both in the presence of spikes throughout the interval, or absence of spikes altogether. To avoid false positives, all flags are removed, and the interval is considered to have no spikes. Cases with many spikes throughout the interval can be filtered based on the intra-hour variability flag.

page 29, step 3: why is it not desirable to also flag short-duration spikes? Couldn’t these originate from a very local source, such as a generator?

This unflagging step concerns data points that we consider statistical outliers. Since step 1 features a variability threshold of 3.5 standard deviations, it flags data points with natural variation. Assuming a Gaussian distribution of the variability and no spikes, 0.05 % of all valid data points would be expected to exceed the threshold of 3.5 standard deviations. Therefore, we consider individual flagged data points, or very small groups thereof, false positives. Therefore, they are unflagged in step 3. In addition to the above reasons, their impact is negligible and they would complicate further steps, which is why they are unflagged in step 3.

page 30, line 2: why choose a threshold of 8 std deviations? this seems arbitrary

All parameters in the spike detection algorithm were tuned to work with the data from Ambarchik, so might seem arbitrary. We do not claim that these settings can be applied to different sites without further review. The chosen criteria worked best for our own site, and we believe they may also provide good starting values in case the procedure is applied to other datasets.

page 31, Figure D.1: This figure shows the utility of using an algorithm to remove spikes and it does seem to work reasonably well for this case. But the complexity of the strategy is concerning. When the data is distributed, it would be best if the flagging for spike-detection is reported separately from other types of flagging (e.g. flagging after transitions, flagging for maintenance) so that the end user can consider alternative strategies.

We fully agree with the reviewer on this topic. Because of the complexity of the algorithm, we distribute hourly data both with and without application of the spike detection algorithm. Note, however, that the impact of the algorithm on atmospheric data was small anyway (see Table 3 in the manuscript).

page 32, E.1 It would be useful to describe the Allan variance of the analyzer and to distinguish

between random error that reduces with averaging versus uncertainties that result from systematic errors that cannot be reduced by averaging.

We add a description of random and systematic uncertainty components together with Allan deviation plots as new section Appendix E.3. To summarize, our error model relies on the following components: uncertainty due the calibration strategy, uncertainty of the water correction, instrument drift and noise. Of these uncertainties, only instrument drift and noise (σ_u, σ_y') are affected by averaging. For a better understanding of this component, we computed the Allan deviation based on the 12-day calibration measurement that was used for estimating σ_u (see revised manuscript). The results (Fig. 2, also included in the revised manuscript) indicate that further averaging does not improve the uncertainty due to the random components, i.e. instrument noise and drift. The Allan deviation estimates of our analyzer are within the range of those for several gas analyzers of the same type as ours, as documented in Yver Kwok et al. (2015).

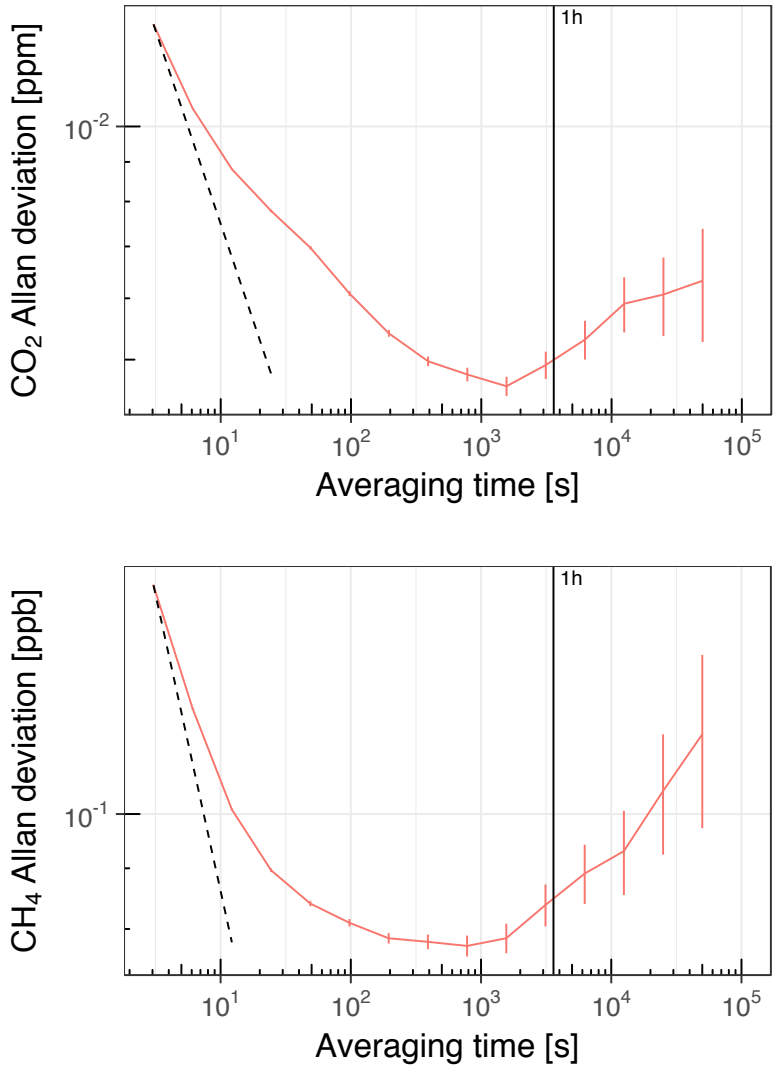

Fig. 2: Allan deviation of the CO2 and CH4 readings of the CRDS analyzer in Ambarchik. Values are based on one 12-day measurement of dry air from a gas tank in the lab prior to field deployment. The averaging time is cut off where the error gets too large for a meaningful interpretation of the result. The vertical line denotes an averaging

time of one hour. The dashed line corresponds to white noise (slope -0.5), scaled to coincide with the first data point of the Allan deviation.

Specifically, if laboratory tests or field calibration data can be used to estimate the random component at the native frequency of the measurement and for hourly averages, then that would allow the user to determine when atmospheric variability exceeds the random noise of the Picarro analyzer. This can help with data selection and weighting in inverse modeling. See the discussion of "sensor precision and atmospheric variability" in the recently released GGMT 2017 meeting report (GAW Report 242).

The analyzer signal drift and precision for hourly averages was reported in Appendix E as σ_u. The values were: 0.013 ppm CO2, 0.25 ppb CH4. We add these values, and a statement that they might be used to distinguish between analyzer signal and atmospheric variability, to Sect. 3.4 "Uncertainty in CO2 and CH4 measurements".

A related question is whether the standard error of the fit takes into account the 120 day smoothing of the coefficients. For a simple case with a uniform (boxcar) 120 day weighting, there would be approximately 24 separate calibration episodes = 70 degrees of freedom. The standard error is substantially reduced compared to a single calibration episode. An example with realistic values and errors is given in the attachment (coded in R) and improvement in the fit coef uncertainties and the overall residual standard error of the fit is evident when multiple calibrations are combined. Here I neglected noise on the assigned values. It should be straightforward to adapt the equations from Andrews et al., 2014 Appendix D to account for the tricubic kernel weighting if that method is demonstrably superior to simple boxcar smoothing.

We thank the reviewer for spotting this error in our analysis! Indeed, our uncertainties did not account for the error reduction achieved by smoothing the coefficients – which was of course the purpose of the smoothing. We update our analysis by calculating new calibration fit functions, this time based on using all calibration episodes in the averaging window and with the weights used for averaging coefficients. We confirmed that our averaged coefficients were virtually identical to those based on these weighted fits, and did therefore not change them. We then recomputed the uncertainty components that were affected. Below, we summarize which terms were affected how by this correction:

$z_{(\alpha,f)}$: Reduced because of the increased degrees of freedom (~1 now).

σ'_y: Increased for CH4 because residuals now correctly account for instrument drift. For CO2, the values decreased because the residuals were typically larger than the drift.

se_{fit}: Competing effects of reduction due to the larger number of observations and increase because of instrument drift.

Also, we previously reported $z_{(\alpha,f)}\sigma'_y$ and $z_{(\alpha,f)}se_{fit}$ in Table E.1, instead of σ'_y and se_{fit}. This was supposed to give a better sense of the contribution of these quantities to the total error. However, since this fact was omitted in the manuscript, and $z_{(\alpha,f)}$ is roughly equal to 1 now, we switch to reporting σ'_y and se_{fit} as stated in the table.

The standard error of the fit (se_{fit}) is now computed based on the equations given in Andrews et al. (2014), but modified for the case with varying weights (following Taylor, 1997).

We update the description of the uncertainty estimation in Appendix E.1, including the formulas for the uncertainty components for fits with weights, the values of the affected components in Table E.1 and the final uncertainty estimates in Fig. E.1.

Note that instrument drift now affects both se_{fit} and σ_u. However, while se_{fit} captures drift on the time scale of the averaging window of 120 days, it treats drift significantly below this time scale as noise. Thus, the contribution of drift on these shorter timescales to se_{fit}would tend toward 0 for larger numbers of measurements. By contrast, our estimate of σ_u is based on measurements over the significantly shorter period of 12 days. Therefore, it serves as an estimate of short-term drift not captured by the smoothed calibration coefficients, and is left unchanged.

And/or you could use the "residual standard error" of the fit to find the optimal averaging window and weighting strategy.

This would indeed provide an interesting study objective. However, it is outside the scope of the presented manuscript, and therefore needs to be postponed to a follow-up study.

The "sigma prime y" term in E.4 will also be affected by analyzer noise, and may be smaller for an hourly average value than for a single calibration episode. In any case, it is important to describe the random error characteristics of the analyzer and the individual calibration episodes.

As described above, we add a section on “Random and systematic uncertainty components” to the appendix (Appendix E.3), where we discuss these considerations. In short, random uncertainty components only played a minor role in our uncertainty estimates.

Please also note the supplement to this comment: https://www.atmos-meas-tech-discuss.net/amt-2018-325/amt-2018-325-RC2- supplement.pdf