Thank you for this detailed and instructive feedback. We have made several changes to the manuscript to address these comments; changes to the manuscript are noted below.

The abstract needs some adjustments to better summarize the main topic of this work, i.e. optimization of the CFA method, intercomparison of CFA-CRDS with discrete sample CRDS measurements, error analysis and noise contribution, etc. The authors should refrain to use strong wording such as "routine measurements", because if that would be the case then their current work would not be relevant. Also the claim of high-precision measurements of Δ17O <5 per meg is questionable. Considering the reported 0.3 cm/min melt rate, the 1.4 cm resolution would correspond to about 270 s measurement time on CRDS that would result in about 25 per meg precision, in best case (see Allan deviation plot). The precision of ~5 per meg can eventually be achieved at averaging times longer than 3000 s, which corresponds to depth averages of ~15 cm. Obviously, more clarification is needed here.

We have updated the abstract to better reflect the conclusions of the paper, as suggested.

Another aspect that requires further discussion is the "calibration errors". First, it needs a clear definition and second, the interpretation of the data should be reconsidered. The authors present a large number (>40x, 3 h, over seven weeks) measurements of the reference waters (SW, SPS2, and CW), but they don't say/show anything related to the CW, although this is supposed to be used as an independent verification of the calibration. Also, the information about the two-point linear calibration curve (offset, slope and their variation across the individual measurements) is not given.

We have clarified the spread of the CW measurements in the manuscript and have added new text about the calibration error, including a new Fig. 4 that provides the calibration slope and intercept data.

Furthermore, a detailed description of the many limitation and critical role of a constant flow across the CFA is presented, but then the calibration gases are used on much lower flows compared to the ice core measurements. It is expected that changes in the flow would have a significant impact on the droplet formation in the vaporizer and that adsorption, memory effects, and isotope fractionation can also be changed. Considering all these details, it is difficult to see any direct and easy link that would allow for the conclusion drawn by the authors about the dominating role of "calibration error" in the CFA-CRDS method.

We have updated the text to clarify that downstream of the melt water vent tee, calibration standards and sample melt are treated identically, thus eliminating concern

about difference in fractionation within the vaporization system. The pressure at PI-1 is maintained at ambient conditions during both reference water and sample melt processing, and downstream flow rates are not adjusted.

Finally, the direct comparison of the CRDS stability (Allan variance) with the CFA reproducibility on Fig 7 is questionable. The Allan variance assumes a continuous stream of data, condition that is not fulfilled by the discrete measurements. Plots of the two-sample variance from different measurements based upon successive recordings of samples display a 'shoulder' effect of the reduced duty cycle on the system performance and the first deviation from the $1/\tau$ slope is a duty cycle effect and not an indicator for an accuracy (calibration) problem.

We believe that Reviewer #1 has misinterpreted our analysis. We do not use the Allan variance to evaluate calibration (i.e., accuracy) – rather, we use the Allan variance to demonstrate the precision of the system, and then compare it with the precision of CFA measurements. We have modified the manuscript text to make the analysis shown in Figure 7 more explicit.

Reviewer #1 suggests that two-sample variance from different measurements will display a "shoulder" effect of the reduced duty cycle, and we agree that the necessary lapse in time between our continuous measurements would affect the variance of the uncalibrated datastream due to, e.g., instrument drift occurring over long timescales between CRDS measurements. However, each of the nine CFA-CRDS timeseries used for this analysis has been calibrated to independent reference water measurements made immediately before the ice core data, and the variance among CFA-CRDS datasets is therefore a function of both the instrumental noise within the integration window of duration τ and also the calibration. The calibration should account for long-term drift in the mean value of the uncalibrated ice core measurements because the calibration standards would also be influenced by the drift. If the calibration is optimal, calibrated data should exhibit a similar relationship between variance and integration time as the Allan variance analysis. Despite this, we find that the precision of our calibrated data is worse than suggested by the Allan variance. We test the impact of the calibration intercept – which affects the magnitude of the mean signal but not the amplitude of the seasonal variability within the core. Shifting the calibration intercepts so that the mean values are equal results in a standard deviation to integration time relationship that mirrors the Allan variance relationship. This is similar to the intercept adjustment made by Steig et al. 2021 and suggests that calibration techniques can be improved with additional reference water information.

Specific comments:

Pg4, l.94. The authors should quantify the "several times more vapor".

We have clarified that the volume of water vapor produced at the vaporizer is approximately thirty times greater than required for analysis.

Pg4, l101. A pressure sensor monitors pressure and not flow conditions.

Agreed -- we have reworded this to be more precise.

Pg4, l111. Is there no issue (adsorption, memory) with having water carried over PFA tubing? Why do not use electro-polished stainless-steel tubing with inert coating instead?

Thank you for this suggestion. Although other materials may reduce adsorption throughout the system, transparent PFA tubing is advantageous for visually inspecting process lines while troubleshooting and was therefore selected for this work. We have added this detail to the manuscript.

Pg5, l130. The expression "analytical cavity" is not correct. I suggest using "optical cavity". Check for all instances across the manuscript. Also the "required sample volume" is not appropriate. The volume of the optical cavity is fixed as well as the pressure at which the CRDS operates.

We have updated this language as suggested.

Pg5, l131. What is meant by "measured volume"?

In some instances this had referred to the vapor that enters the optical cavity, and in others it referred to the liquid that enters the vaporizer; we rephrased this throughout for clarity.

Pg5, l134. Replace "all system instrumentation" by "sample handling system"

We have updated this language as suggested.

Pg5, l137. Consider simplifying the wording and replace "system instrumentation volume and system tubing diameter were minimized" with "the overall sample handling system volume was minimized"

We have updated this language as suggested for clarity.

Pg5, l140. Quantify the "excess liquid". How much compared to the measured sample?

Approximately 6x more water volume is vented and containerized than is vaporized. We have clarified this in the text.

Pg5, l143. Quantify "several times more"

We have clarified this in the text – approximately 6x more water volume and 30x more vapor volume is produced than is analyzed.

Pg.5, l145. The pump rate is minimized when measuring reference water. The authors should specify by how much is the flow reduced and comment on the expected effects due to the change in flow especially in terms of droplet formation in the vaporizer, adsorption, and isotope fractionation (see also my general comments).

While measuring reference waters, the PUMP-1 rate is reduced to match the flow rate of PUMP-2 (measured at 26uL/min). While measuring ice core melt, the PUMP-1 rate is set to accommodate the ice core melt stream (~450uL/min between sample and waste lines combined). The liquid line pressure at PI-1 and the flow rate setting at PUMP-2 are both held constant between sample and reference water measurements, so we do not expect or observe any flow changes at the vaporizer (the rate is ~26uL/min for both reference water and sample melt). We modified language about this in the manuscript for clarity.

Pg.7, l198. The anticorrelation between water vapor and δ18O is explained by two different effects: 1) incomplete vaporization, and 2) insufficient backpressure. The authors should be more consistent and clearly state which one applies. Is the correlation a real physical effect, due to e.g. isotope fractionation, mixing, etc., or is it simply reflecting the δ18O dependence on water vapor amount fraction of the CRDS?

We have reworded this section to clarify. We have observed that insufficient backpressure on the peristaltic pump causes larger, intermittent droplets to form within the vaporizer (instead of smaller, higher frequency droplets). These larger droplets do not completely vaporize within the vaporizer tee. The anticorrelation is a physical effect due to isotope fractionation within the vaporizer.

Pg.7, l201. What does the "apparent fractionation" means? An instrumental response?

We have updated this language to say “observed fractionation” – the vaporizer can be manipulated to cause fractionation within the chamber (by overwhelming the vaporizer with large droplets that do not vaporize immediately). The incomplete vaporization

response is characteristic and predictable. We have also pointed out this response with labels in Figure 3.

Pg.7, l205. The authors associate flow inconsistencies with isotope fractionation, but it is a misleading argument, because without excluding artifacts from the CRDS measurements itself it is difficult to disentangle the observed effects. In other words, the authors should discuss the mechanism behind the isotope fractionation generated by the variations in the flow.

We have reworded this text to better explain the relationship between the flow inconsistencies into the vaporizer and the observed fractionation.

Pg.7, l205. There is a list of many significant interventions: adjusting the peristaltic pump rate, replacing filter screens, adjusting FV-1, replacing peristaltic pump tubing, replacing or cleaning the capillary tube, or cleaning the vaporizer. The authors should give a more detailed discussion how often are these interventions necessary, what does it mean in terms of operation down-time and what is the impact on the calibration scale. Changing so many items should definitely result in rather different system response in terms of memory and surface effects, etc.

System changes are made before ice core measurements or between reference water measurements as needed. Filter screens were occasionally replaced during analysis, which caused brief flow disturbances at the vaporizer but did not otherwise impact analysis. This has been clarified in the text.

Pg.7, l210. The authors should explain how they are able to perform the automatic measurements while routinely tune the system to maintain steady pressure.

This language has been modified for clarity. We typically have considered the ongoing sequence of automated reference waters to be a “continuous” operation, but this is not the same as a continuous ice core measurement. The system is tuned before reference water measurements that are associated with ice core samples, or at other times in between reference water measurements when the system was observed to be out of balance. These adjustments were very intermittent because operators were rarely in the physical lab space aside from ice core measurements due to COVID-19.

Pg.8, l221. What is a manual observation?

We have clarified that this was a visual (and not electronic) observation of core height.

Pg.8, l244. What is the amount of rejected data compared to the entire set of measurements?

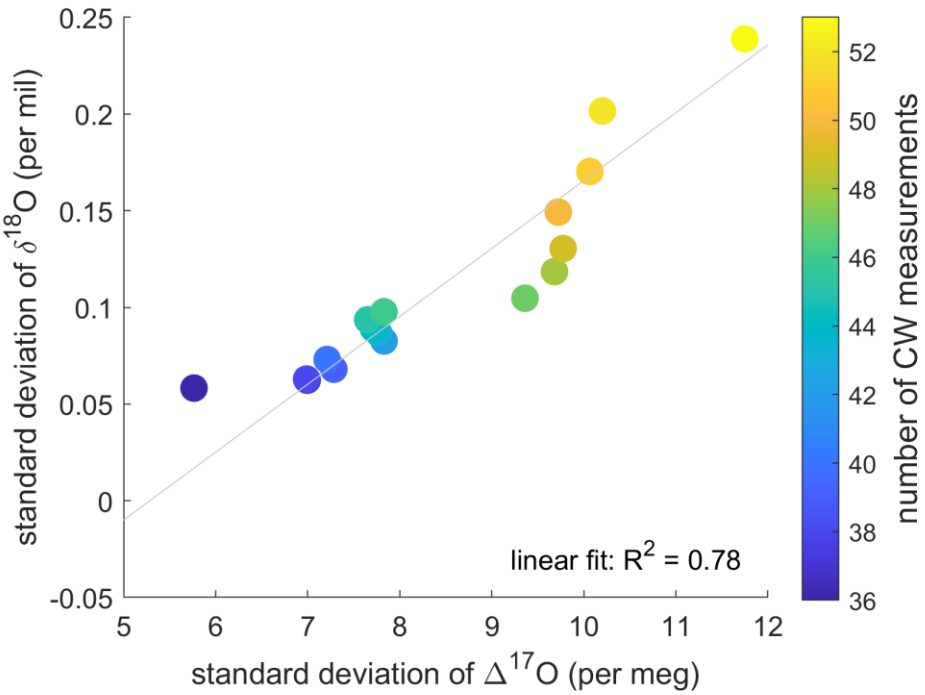
Due to COVID-19 lab restrictions, the reference waters were typically measured without oversight (reducing quality of some measurements if the system became imbalanced) and there were a few lengthy periods of downtime during the analysis window (when no operator was available for several days to reset the system). The measurements presented here represent approximately 50% of the analysis time. All measurements that were made with operator oversight are included in the final dataset.

Pg.9, l250. It would be very helpful to provide the scatter of the 47 individual 3 h measurements of the reference water to illustrate the stability of the CFA-CRDS system for this fundamental step.

We updated the text to state that the mean of all CW measurements is 25 and that the standard deviation is <12 per meg (n=53). For the most tightly clustered measurements of d18O, the mean of these CW measurements is 25 and the standard deviation is <6 per meg (n=36).

We also include here a comparison of CW variability for d18O and D17O to highlight the relationship between these calibrated datasets. To generate this figure, we took the standard deviation for d18O and D17O measurements of CW that had been filtered by progressively more restrictive variability thresholds for d18O. All measurements (n=53) have a standard deviation <0.25 per mil for d18O and <12 per meg for D17O, but we also know that some of these measurements were taken when vaporization conditions were not optimal (see Figure 3 for optimal conditions). Here we confirm that fractionation observed in d18O has a direct relationship to the errors in D17O – it also shows that all CW measurements are acceptably precise, but the best measurements

(as indicated by d18O variability) are also excellent for D17O (<6 per meg).



Pg.9, l264. This sentence is misleading. I suggest to modify it, e.g. "our ice core measurements cover of about two years period"

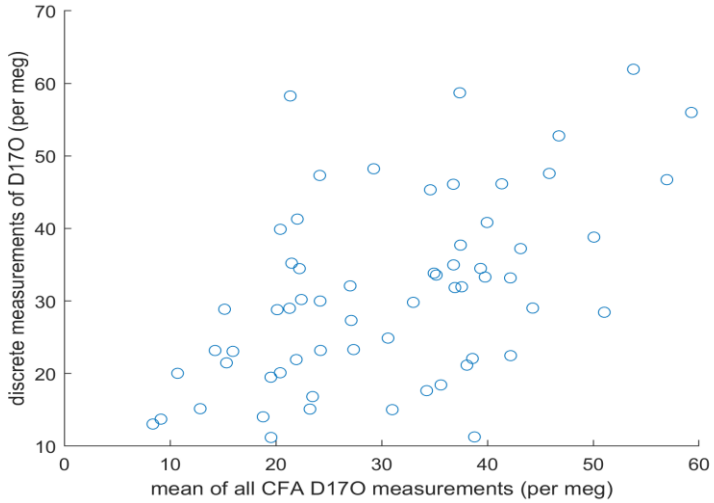
We have modified the text as suggested.

Pg.9, l273. In contrast to the author's statement, a correlation coefficient of 0.52 could either be interpreted as a “good” or “moderate” correlation, depending on the applied rule of thumb. The observed correlation may not necessarily be a good estimate for the population correlation coefficient, because samples are inevitably affected by chance. Therefore, the observed coefficient should always be accompanied by a confidence interval (95 %), which provides the range of plausible values of the coefficient in the population from which the data were sampled. Furthermore, the correlation coefficient of 0.52 corresponds to a coefficient of determination (R^2) of 0.27, suggesting that only about 27 % of the variability can be “explained”. Finally, since both data are observations, a Pearson correlation analysis would be more appropriate here. In this case, both variables are assumed to be subject to natural random variation.

The Pearson correlation analysis was used to determine the correlation coefficient. The r value for the 1.4-cm data is significant with 99% confidence; this has been added to the text.

Pg.10, l274. Again, this statement is inappropriate since the deviations are comparable with the seasonal variation. It would be instructive considering a scatter plot using the CFA and discrete CRDS data shown in Fig7 (bottom).

Because the average CFA dataset includes nine measurements, the total analysis time associated with each depth interval is greater than 2000 s, and the noise expected (and observed) is substantially smaller than the seasonal variations. We have clarified this in the text, and include here the scatter plot that is suggested to show the correlation between CFA and discrete measurements.



Pg.10, l185. Although, this in principle holds, the slower ice melting would result in slow response time across the CFA, which would then have an impact on the achievable resolution. In general, instead of hypothesizing what would in principle be possible, the authors should consider only those cases that are realistic for high precision and routine ice core measurements.

We agree; we have removed this section from the text.

Pg.10, l289. The term of high-frequency instrumental noise is misleading. Use 1 Hz precision instead.

We have reworded this as suggested.

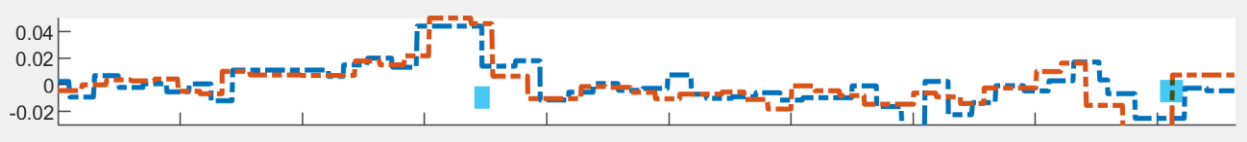
Pg.10, l292. The authors should explain what they mean under calibration error and what are the disproportionate drifts in d17O and d18O. Is there any systematic investigation of the oxygen isotope fractionation during vaporization? If yes, it would be

helpful for the reader to know its magnitude, reproducibility and dependency on various factors, such as flow, pressure, etc. Without these facts the claim cannot be proven.

We have reworded this section for clarity and have also added a new section and new Fig. 4, which clarify the calibration process and the effects of fractionation on the calibration values. We have also added text throughout the document to better describe calibration procedures and related errors.

In addition to Fig. 4, we below include a plot that examines changes in calibration error over time. Earlier work has identified that calibration offsets can impact the mean values of CFA-CRDS D17O data, so we are not surprised to find that over time, we observe similar calibration changes in our data. However, we do try to track these changes and use them to tailor maintenance and operating schedules, which we added to the discussion section of the manuscript. In the absence of fractionation, the calibration slopes and intercepts for d17O and d18O would be perfectly correlated. Monitoring the differences between calibration values for d17O and those for d18O therefore provides a way to directly assess changes in fractionation through time.

Here, we find the residual of m_d17O compared to the mean relationship between m_d17O and m_d18O, which has an R^2 value of 0.99 (blue line in the plot below is the residual of m_d17O). We do the same for values of b, which have an R^2 of 0.98 (orange line in the plot below). Despite very high correlations (and calibration differences that are completely inconsequential for d17O and d18O values), these very small deviations in d17O and d18O calibrations document fractionation that affects the calibration of D17O over time. In the plot below, the blue squares mark two times when the vaporizer fittings were cleaned. We have added some commentary about using these tools for method and calibration strategy development to the manuscript.



(y axis is in per mil for b (unitless for m), x shows the total duration of all measurements (same as Fig. 4)). Slope-derived residuals in blue, intercept-derived residuals in orange.

Pg.10, l315. The Allan variance analysis doesn't determine the theoretical variability, but the observed one.

We have clarified this wording as suggested.

Pg.10, l316. Define "internal noise".

We have corrected this to say “signal noise.”

Pg.11, l320. This statement is not clear enough. If the reference water is handled in the same way as the melt water then how is it possible that its measurement does not account for the variability from the CFA system?

We agree that this statement was confusing as written; we have reworded this section for clarity. The Allan variance assesses the signal stability of a single continuous stream of data, but we are interested in how the precision during the measurement window compares with the precision of calibrated measurements on longer timescales. The repeat CFA-CRDS measurements allow us to compare these two things because we can assess the variability of individual measurements about the mean value at any given depth in our core.

Pg.12, l341. Again, the 1.39 cm resolution would correspond to 270 s measurement time leading to about 25 per meg precision (1σ) according to the Allan variance. Thus, the signal-to-noise on the seasonal cycle is less than 2.

Each depth interval includes 2000+ seconds of data and we have shown that the precision of this dataset it substantially smaller than the seasonal cycle. We have clarified this throughout the text where combined CFA-CRDS measurement data are discussed.

Fig.7: The Allan variance plot has a remarkable character. It seems that the CRDS stability is extraordinary as the deviation continues to decrease even after 10^4 second (2.8 h). The authors should comment on why do they stop at this stage and consider the 2 per meg as best precision. It would be very interesting to see how long this continues. The authors should perform even longer reference water measurements to explore the limits where drifts start to dominate. At such instrumental stability, there is no need to make any calibrations for hours. This would imply that a 50 cm ice core could be easily measured in one run. Why this is not shown in this work?

We agree that the signal is remarkably stable on these timescales. In this paper, we include an Allan variance for an 8.5-h measurement period of Seattle water because this is the longest run of a single reference water that was made during the ice core measurement period, and it is therefore the most representative of the system configuration that was used for this analysis.

There is a "bump" at 10 s averaging time. Normally, this indicate a periodic oscillation in the system. Can the authors comment on the origin of this deviation from the white-noise at that time-scale?

We have observed this feature in all data from the instrument and are confident that it is an instrumental effect and not related to our sample handling system. It is documented in Schauer et al., 2014 and in Steig et al., 2021. While we do not fully understand the cause of this feature, we do not use data on such short timecales and therefore this feature does not affect our analysis.

Technical comments:

Abstract l.8-10: I recommend combining the two sentences and remove the "recent advances", because the CRDS technology for oxygen isotope measurements is almost 10 years old without any significant development since then. My suggestion: "... continuous-flow analysis (CFA) methods coupled to CRDS instrument allow for simultaneous measurements ...."

We have revised this statement as suggested.

I also suggest deleting the last sentence in the abstract.

We have removed the last sentence as suggested.

Pg2,l.30. Replace "routine" by "were demonstrated" since many of the cited references give a demonstration and show the feasibility of the technology rather than presenting routine ice core measurements.

We have reworded this as suggested.

Pg2, l.32. Replace "laser spectroscopy instrument" with "a laser spectrometer"

We have revised this as suggested throughout the manuscript.

Pg3, l.67. I suggest to write "CRDS spectrometer (L2140-I, Picarro Inc.)"

We have revised this as suggested.

Pg3, l.84. Remove space between value and degree, i.e. 200°C (look for other instances in the manuscript as well)

We have corrected this spacing here and throughout the manuscript.

Pg4, l.98. Replace "near-instantaneous analyzer output" by "CRDS values". The authors can assume that the reader is familiar with the CRDS and knows that this instruments report the results on a 1 Hz rate. I suggest removing all instances of "instantaneous".

We have revised this as suggested, here and in other instances throughout the manuscript.

Pg4, l102. I suggest replacing the very vague sentence "Instantaneous instrument output that reflects the internal vaporizer conditions and monitoring of the CFA line pressures provide information that is used..." by "These information is used..."

We have reworded this as suggested.

Pg4, l107. Replace the "x" by "×". For sake of simplicity write 30×30 mm^2

We have changed this as suggested.

Pg5, l140. Change "ice-core" to "ice core". Check for all instances across the manuscript.

We have updated this throughout the text.