The replies below are structured to comply with the suggested structure of the “Author’s response” with (1) referee comment (2) author comment (3) implemented revisions. Referee/editor’s comments are in blue-italic and authors’ response in black font. Implemented changes are highlighted in red font.

Response to reviewer #2.

The authors thank again reviewer #2 for providing new suggestions and for being rigorous with this review - this is highly appreciated and helped us improve the manuscript.

I experienced various HydroC sensors that not only show large discrepancies from established methods (tens of nanomoles per liter) but also randomly distributed time series patterns during field deployments in the past, so it’s good to read that the new TDLAS-based sensors seem to be more reliable. Hopefully, the statement "post/intermittent calibration was neither recommended by the manufacturer nor found necessary" is not entirely serious... Verification of scientific data is more than essential, even if a manufacturer states that continuous verification and/or calibration measurements are not necessary!

So, in general, it remained quite difficult for me to come to a decision here due to the lack of any quality control for the CH4 concentration data obtained by the two HydroC CH4 sensors (missing post/intermittent calibration or verification during a one-year deployment of in situ sensors!). However, I agree that this study found significantly larger fluctuations in the CH4 concentrations than can be expected from the sensor and, in addition, this study is rather focusing on relative values (methane variability and its driving factors) than on absolute values (calculation of methane fluxes and inventories). Therefore, I have decided to accept most changes in the manuscript provided by the authors - but I still have the following requests for minor changes in the manuscript:

• The material & methods section should be extended by a sentence indicating that no post and/or intermittent calibration or validation of the sensors was performed. In addition, a further section in the manuscript should mention that the deployed sensors do have strong limitations for providing quantitative measurement, but that, however, the obtained CH4 data should be sufficient to support the scientific results in this specific manuscript (e.g. something similar to your statement „Even though modern HydroC CH4s with our pump/membrane/detector setup can give decent accuracy and be applicable in a wide range of settings, it might still be relevant to acknowledge that the analysis and discussions in the current manuscript mainly concerns large changes and high concentrations. Considering this, we believe the response time corrected Contros HydroC CH4 data should be more than sufficient to support the scientific results and inferences described in our manuscript“).

The HydroC CH_4 sensor deployed in our study is a different sensor in several crucial aspects compared to the older NDIR systems. These changes apply directly to the ability of the sensor to provide quantitative measurements and long-term stability. In our opinion, it is misleading to state that the sensor has strong limitations in providing quantitative measurements as long as the long response time due to the membrane exchange process is corrected for (which is the case in the present study, using the method detailed in Dølven et al., 2021). This procedure also lets us explicitly report the data accuracy (Fig. B1b), something which was in practice impossible in the past due to the long response time, thereby also conveying explicit probabilistic (confidence intervals) information on the limitations of the data we present.

Evaluation of a measurement should be related to the properties of the phenomenon of interest – some applications require very high accuracy, while others require fast response times, etc. Since we are confident that the data quality is more than sufficient to support our inferences, we find that descriptions of sensor performance should be anchored in reporting our estimated measurement uncertainties, intervals, and general principles etc, rather than stating that the sensor has strong limitations for providing quantitative measurements (for certain applications it probably has - most sensors have applications where they fall short).

That being said, we completely agree with the reviewer that verification of scientific data is essential and long-term deployments are especially challenging – this applies to all deployed sensors. Nonetheless, some sensors are more prone to drifts and erratic behavior than others and systems based on NDIR have a wide range of challenges, especially in long-term deployments with varying environmental parameters. But many of these challenges are alone overcome by using a TDLAS detector (I lack a full account of the various HydroC revisions, but I know other improvements have also been implemented, such as the pump system). Below I elaborate on some of the crucial differences between the TDLAS based system and a NDIR based system in long-term deployments which can hopefully clarify some aspects and justify why we find the addition of the suggested statements (on poor long-term stability and inability to provide quantitative measurements) inappropriate to add without further explanations (implemented changes are found in red below).

For reference see e.g. Lackner (2007), Shemshad et al. (2012), Wang et al. (2019), or Zhang et al. (2021), but also technical description of the sensor https://www.4h-jena.de/en/maritime-technologies/sensors/hydrocrch4/.

Quantitative quality of the measurements

As opposed to NDIR, which emits light at a comparatively wide and constant (assuming no drift) frequency band constrained using an interference filter, a TDLAS based system performs its measurement by scanning over the absorption line(s) of the gas of interest using a laser (i.e. very narrowband, near monochromatic, near coherent light) source with tunable output wavelength. This leads to several advantages with regards to quantitative quality over NDIR:

1. Lower detection limit due to a specifically tuned light source (NDIR also selects appropriate wavelengths but uses an interference filter which usually has a much cruder spectral resolution.)
2. Much less susceptibility to pressure broadening effects (in NDIR the broad band results in pressure modulation of the signal due to changes in the absorption line surface)

3. Overlapping absorption bands are a challenge with NDIR technology because of the constant and wider (and non-discriminatory) frequency band. This is not an issue with TDLAS due to the very narrow and tunable frequency band which makes it possible to explicitly identify where absorption bands of different gases interlace and thus obtain a clear separation between different gases.

All these aspects increase the quantitative quality of the data.

Long-term stability/calibration

NDIR systems are often challenged by poor long-term stability due to several effects such as variations/degradations in the light source over time (both causing spectral and intensity shifts), degradation of detector and contaminations in the measuring chamber. This makes intermittent calibration by e.g. zero gas crucial to monitor and compensate for long-term stability in NDIR based systems (such as the zeroing technique in the Contros HydroC CO_2, see Fitzek, 2014). This is not necessary in the same way for TDLAS based systems because the frequency (spectral band) of the light source (laser) is controlled, which not only removes the issue of drift in the spectral band, but also makes it possible to compensate/check for drift by tuning the laser between non-absorbing and CH_4 absorbing wavelengths. The TDLAS detector in the Contros HydroC CH_4 does this. Although we agree that intermittent calibration would be advantageous for all sensors in such a long-term deployment, this is not specifically necessary for the TDLAS HydroC system (but it would be, for a NDIR-based system).

The main challenge with quantitative quality provided by TDLAS based sensors deployed in dynamic domains is their long response time, which alone can easily result in large discrepancies with other methods and result in raw data which does not represent reality. However, we correct for this by doing a response time correction of the signal. Furthermore, we found no suspicious behavior in neither data nor meta data (i.e. internal temperature, pressure, etc.) during post-processing or in the evaluation of model fit residuals in the response time correction procedure (which modulated the signal variance and was otherwise Gaussian - as expected). All this considered, we believe that the quantitative quality of the measurements should be more than sufficient in our application. We also cannot find concrete evidence to support, nor a relevant reason to add a sentence stating that that the sensor has strong limitations when it comes to providing quantitative measurements in the methods section of this manuscript.

Nonetheless, we agree with the reviewer that it is important to clarify these aspects in the method section and have attempted to incorporate the suggested changes ("*The material & methods section should be extended by a sentence indicating that no post and/or intermittent calibration or validation of the sensors was performed. In addition, a further section in the manuscript should mention that the deployed sensors do have strong limitations for providing quantitative measurement*"), but at the same time reflect what we iterated above by rewriting/adding the following text in the method section (ln 96-104):

Uncertainty ranges for the CH_4 sensor data are reported as 95% confidence intervals and typically vary between 5 and 20% (Figure B1b). We did no post and/or intermittent validation. Although always an advantage for all sensors in long-term deployments, this is not a requirement for the TDLAS based sensor (as opposed to NDIR), due to its high long-term stability. Standard post-processing (e.g. inspection of meta data such as internal pressure and temperature) and evaluation of fit residuals in the response time correction procedure (see Appendix B and Dølven et al. (2021)) also indicated consistent sensor behavior. It is also worth noting that the current manuscript concerns large changes and high concentrations and we are confident that the quality of the response time corrected Contros HydroC CH4 is sufficient to support the inferences described herein.

We also added a description of the error fit residuals in Appendix B (ln 416-419):

Inspection of model fit residuals showed a slight modulation following the variance in the signal, explained by our choice to use the same 3-minute measurement grid across a relatively wide variance ranges, but were otherwise Gaussian. Although expected, this indicates that errors might be slightly overestimated for low-variance sections of the time-series and vice versa for high-variance sections.

And clarified which version of the sensor we deployed (ln 85-88):

The deployed HydroC CH4, being a younger iteration of the sensor, rely on a Tunable Diode Laser Absorption Spectrometry (TDLAS) detector (rather than non-dispersive infrared spectrometry (NDIR)), while the CO2 sensors use NDIR detectors.

• The authors did not follow my suggestion to implement some basic calculations of CH4 inventories but, however, they came up with an alternative approach to emphasize the urgent need of continuous measurements to detect the temporal variability of seabed seepages. I like the idea of calculating the expected error from unresolved short-term variability for a hypothetical discrete water sampling survey seeking to estimate seep site averages, the newly added section (line 204-242) and its corresponding figure (Figure 4). Note: Correct the typo in the y-axis label and consider using a colorblind-friendly palette

Thank you for acknowledging this and for suggesting a deeper analysis on what these measurements mean for budget estimates. We fixed the typo and changed the color palette which should now be colorblind friendly.

• There are still various fragments of grids and box lines of the individual subfigures (mostly grey lines) in Figure 2, but I assume (and hope) this will be corrected before publication. Also accounts for Figure 6.

This should now be fixed. The problem apparently arose in certain pdf readers and was dependent on the zoom level on the screen.

Several dois are still not correct formatted, e.g. Silyakova et al. (2020): "https://doi.org/https://doi.org/10.1016/j.csr.2019.104030" – remove the duplicate "https://doi.org/" and also check all other references!

The duplicates were removed and the reference list has been checked. I apologize for not noticing this error in the first round of review where it was already commented on.

Response to editor

Good to hear. We added a short paragraph regarding the calibration/validation of our HydroC CH4 sensor in the methods section - see reply to review above regarding this issue.

L23: We changed y to yr

L35: There was indeed an issue with this sentence so we rewrote it.

L36: we modified the sentence as suggested

L74-L75: The references used here were indeed not consistent with each other. We changed the reference and now use more "classical" values for the region. These values are all in practical salinity.

Table1 & Table 2: we added a line including units below the parameters.

L185: Done

L252: We agree that the use of z-score might be confusing, so we rewrote this part of the text (in addition to in the figure).

L268: Done

L287-288: We removed all water masses abbreviations in the text to make it easier to read.

L293: We removed all abbreviations for Prins Karls Forland

L301: Done

L487: I added the page number and doi to this reference. Was this what was missing?

L500: We added a link to the data-sheet pdf (same iteration of the sensor that we used), since the manual is not available.

L503: Corrected

L510: This manuscript is still under open discussion. We received one favorable review (https://doi.org/10.5194/gi-2021-28-RC1), but are still missing one review.

Other changes

We changed the ordering of some of the paragraphs in the methods section to better fit the addition above and made minor changes at a few places during a final read through. All is shown in the marked-up version.

References

Dølven, K. O., Vierinen, J., Grilli, R., Triest, J., and Ferré, B.: Response time correction of slow response sensor data by deconvolution of the growth-law equation*, Geoscientific Instrumentation, Methods and Data Systems Discussions [preprint]*, 1–22, https://doi.org/10.5194/gi-2021-28, in review, 2021.

Shemshad, J., Aminossadati, S. M., Kizil, M. S.,: A review of developments in near infrared methane detection based on tunable diode laser, *Sensors and Actuators B: Chemical*, 171-172, 77-92, 2012. *https://doi.org/10.1016/j.snb.2012.06.018*

Wang, F., Jia, S., Wang, Y. and Tang, Z.: Recent Developments in Modulation Spectroscopy for Methane Detection Based on Tunable Diode Laser, *Applied Sciences*, 9(14), 2816, doi:10.3390/app9142816, 2019.

Schiff, H.I., Mackay, G.I. & Bechara, J. The use of tunable diode laser absorption spectroscopy for atmospheric measurements. *Res Chem Intermed*, 20, 525–556, https://doi.org/10.1163/156856794X00441, 1994.

Lackner, Maximilian. Tunable diode laser spectroscopy in the process industries: a review. *Reviews in chemical engineering*. 23. 65, https://doi.org/10.1515/REVCE.2007.23.2.65, 2007.

Zhang, Z., Li, M., Guo, J., Du, B. and Zheng, R.: A Portable Tunable Diode Laser Absorption Spectroscopy System for Dissolved CO_2 Detection Using a High-Efficiency Headspace Equilibrator. *Sensors (Basel, Switzerland)*, *21*(5), 1723, https://doi.org/10.3390/s21051723, 2021.

Fietzek, P., Fiedler, B., Steinhoff, T., & Körtzinger, A.: In situ Quality Assessment of a Novel Underwater pCO2 Sensor Based on Membrane Equilibration and NDIR Spectrometry, *Journal of Atmospheric and Oceanic Technology*, 31(1), 181-196. ,2014.