We are grateful to the referee for this stimulating feedback that brought important additions to our work. Our answers to the comments and questions are provided in red below.

Review of Dogniaux et al. (submitted to AMT)

This group of authors shows their expertise in the method of estimating emission rates from two imagers, Sentinel-2 and Landsat 8. This is a well-written paper. I have no qualms regarding the use English. I believe the paper should be rejected because of the two main criticisms (immediately below), combined with the fact that the methane emission uncertainties are already very large. If the uncertainties were not so large, I think the paper is sufficiently interesting and meritorious to be published.

We agree that uncertainties are large indeed. However, they are actually part of our motivation to submit this work to AMT.

Our work is related to rather novel Earth-imager data exploitation techniques that are easy to implement, already brought very relevant scientific results, and are currently gaining significant momentum within the greenhouse gas remote sensing community. When the Nord Stream leak happened, these techniques or closely-related ones were very swiftly applied to process L8 and S-2B observations which resulted in very rapid communications (on the same day by the International Methane Emissions Observatory: https://twitter.com/MethaneData/status/1575610350548164608) and a Short Communication submitted to "Environmental Science and Ecotechnology" two weeks later (Jia et al, 2022). These communications and other (for example a poster at the AGU 2022 Fall Meeting) claimed that a methane plume had been detected by Landsat 8 and Sentinel-2B, and Jia et al (2022) even provided an emission estimate for the Sentinel-2B observation of 72 ± 38 t/hr, while also acknowledging significant and not so well-defined "uncertainties" ("methodological drawback" may be a more appropriate expression) related to the reflectance of bubbles and part of the plume missing from the observation.

In this submission to AMT, we exactly explain what aspects of the novel techniques are challenged by the Nord Stream 2 observation case, and include a comprehensive uncertainty analysis (improved following this review) that strongly nuances what can be stated about the methane leak based on L8 and S-2B. Therefore, our study can serve as a methodological cautionary tale and detailed discussion of these novel techniques. This is especially important as results from the methane community are increasingly used by a growing group of stakeholders including national governments and international NGOs. We therefore partly revised our Abstract, Sect. 3 and Conclusions to better underline the significance of our work as providing important insight on often used methods.

New text: line 20 – 26	**Our comprehensive uncertainty analysis yields large methane leak rate uncertainty ranges that include zero, with a best estimate of 501±521 t/hr. Thus, no firm conclusion can be drawn from the single or combined overpasses of L8 and S-2B. Within all our Monte Carlo ensembles, positive methane leak rates have higher probabilities (79 – 88%) than negative ones (12 – 21%), thus indicating that L8 and S-2B likely captured a methane-related signal. Overall, we see our work both as a nuanced analysis of L8 and S-2B contributions to quantifying the NS2 leak emissions**

	and as a methodological cautionary tale that builds insight on MBSP and IME sensitivities.

Main criticisms:

1) the ensemble approach for estimating the methane emission rate does not cover the correct range of input values. An ensemble should not span ±1sigma because this only covers 68% of the data. I think it would be more appropriate and simpler to only include +1.0 sigma and -1.0 sigma in the ensemble (and not 0 sigma, or other intermediate increments). Consequently, the authors are underestimating most of their uncertainty sources.

We recognize that the "grid-based" approach we used to build the ensemble falls short in grasping the impact of the full variability of our input parameters. Consequently, we revised our approach and set up a Monte Carlo ensemble that follows best-estimate distributions for all of our six input parameters. The revised uncertainty estimation approach is described in Sect. 2.4 of the revised manuscript. Resulting uncertainties are now ~40% higher in standard deviation, thus confirming that the uncertainties were underestimated However, as explained above, we do think that the size of these uncertainties underscore the scientific significance of our work. The updated uncertainty values are included in section 3.

2) The wind speed calibration coefficients should not have an uncertainty of 5%. I also don't see any justification for such a small value. Did it come from 1.88 versus 2 on L172, which is a 6% difference?

Within the scope of our Monte Carlo ensemble approach, we now use the actual distribution of fit – data mismatches (standard deviation of 1.1 m/s) to perturbate the calculated effective wind speed. This 1.1 m/s standard deviations amounts to 12% and 10% of the average effective wind speed that we compute for L8 and S-2B overpasses. This distribution is given in an additional figure included in the Supplements (Fig. 13).

New text: line 231 – 233	**(6) We account for effective wind speed calibration errors by randomly sampling data – fit mismatch values from the distribution shown in the Supplements (Figure S8). By doing so, we implicitly follow the slightly non-Gaussian skewed distribution that these mismatches show.**

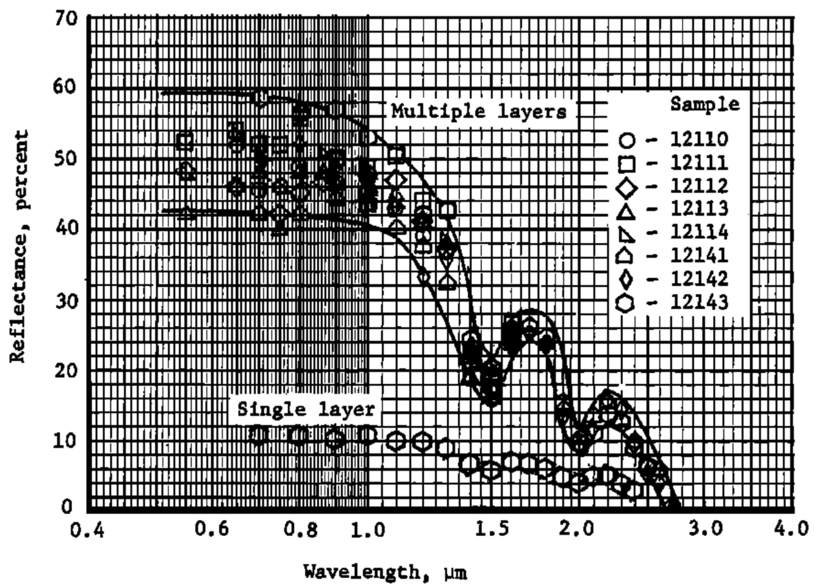
Fig. 1: Please add the dates to the caption for the two reflectance images. Latitude and longitude tick marks would be of interest to the readers.
We added central latitude and longitude ticks, and dates in all satellite image panels included in the paper. The captions have been updated as well and include dates in the revised manuscript.

L107: The linear calibration coefficient varies strongly between Landsat and Sentinel-2. I wonder if these studies are even relevant? Was the reflectance from the bubble monolayer or the multiple-bubble layer used in the Whitlock et al. study? Do they give the same ratio? I could not access the Koepke paper (but the reference is correct). The authors should note that this reflectance ratio should be roughly the same from space and at the ground because

the atmosphere is optically thin. On second thought, is (background) methane a strong enough absorber to affect the ratio (satellite versus ground-based)?
Whitlock et al. used a ground-based radiometer.

These studies were the only reference that we could find to benchmark our space-based results using L8 and S-2B satellite observations. The PDFs that we could find are digitized copies of on-paper versions, and plot quality at the time did not help reading exactly their results. We could evaluate through graphical reading (as written in the original version) that between 1.6 µm and 2.2 µm, the ratio in spectral reflectance should be about 2 or a little lower. Please find below screenshots of the graphs we used.

Whitlock et al, 1982 (we read the multiple layers curve)
https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/GL009i006p00719



(a) Reflectance of foam patches.

Keopke et al, 1984
https://www.researchgate.net/publication/243574607_Effective_reflectance_of_oceanic_whitecps

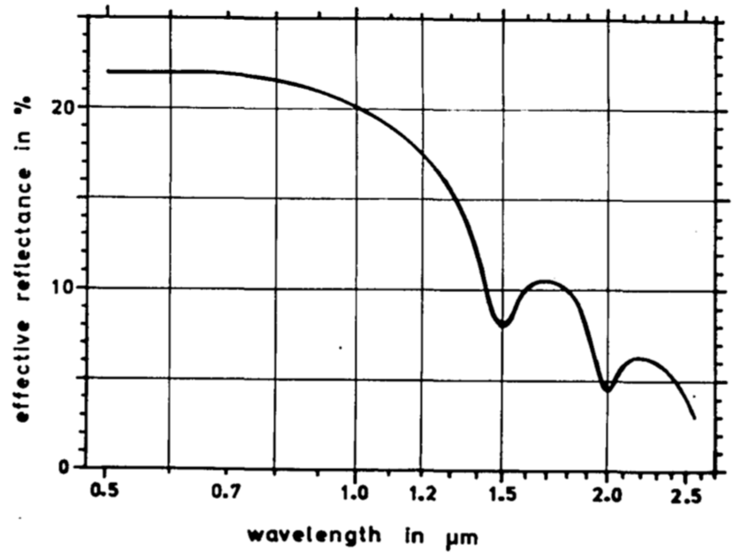


Fig. 8. Effective reflectance of oceanic foam as a function of wavelength, after Eq. (9).

The spectral dependence of reflectance that we assess between these satellite bands are rather the spectral dependence of Top-of-the-atmosphere reflectance, or 'space-effective', because the atmosphere is not optically thin enough to transmit light seamlessly to its top. As shown in Varon et al (2021), these 1.6 and 2.2 μm bands also exhibit spectral lines of water vapor for instance. On the top of these lines, continuum absorption of water vapor, which varies by about one order of magnitude between these two bands, must be added (Shine et al, 2016), and we have seen for other imagers impacts of image-wide water vapor gradients on band ratios that are used to retrieve methane enhancements. The scatter in fitted s1/s2 ratio between ship observations for similar satellites might be explained by water vapor variability for instance. As hypothesized in the comment, methane background concentration variability might also play a role. Other causes could be bi-directional reflectance effects due to slightly different viewing and sun angle geometries (we focused our search around the North and Baltic Seas and in the month preceding and following the NS2 leak to minimize these effects).

Differences between the calibrations obtained for L8 and S-2B could be explained by the small – but existing – differences in the definition intervals of the bands they observe, as well as by their narrow band filter shapes (Zhang et al, 2018). Finally, fewer cases could be observed for L8 because its 30x30 m^2 resolution does not allow to observe as many satisfying ship trails as for S-2B which has a 20x20 m^2 resolution.

All these aspects may explain the variability of s1/s2 ratios obtained for each satellite, and the slight difference in satellite-averaged results. Their precise discussion is however outside the scope of this study.

We added additional explanations on the TOA and ground-based difference between our work and the references in the revised manuscript.

New text: line 138 – 139	[…] We obtain $\bar{c}$ = 1.96 ± 0.23 and $\bar{c}$ = 1.91 ± 0.22 for L8 and S-2B, respectively. **These Top-of-the-Atmosphere reflectance ratios are overall** consistent with results presented by Whitlock et al. (1982) and Koepke (1984) **that were measured on the ground.** Comparing the S-2B result to the slightly higher standard MBSP […]

L132: “through” seems incorrect and terse. I suggest “via the use of”

We corrected this text as suggested.

New text: line 143	[…] reduces $\Delta R = (cs2 - s1)/s1$, which translates to an increase of methane enhancement **via the use of MBSP**.

L151: “mask” could be deleted for simplicity

We delete “mask” in the revised manuscript.

New text: line 169	[…] the plume length, XCH4i, the total column methane enhancement of the i-th **plume pixel**, and ai, the area of this pixel.

L172: This linear regression equation has a much different slope than the one in Fig. 4 of Varon et al. (2018). Was log(U10) also tried? There should be more discussion of why the effective wind speed should exceed U10 for this foamy setting. I think a 5% uncertainty (L197) on the effective wind speed relationship is a gross underestimate. Given the effective wind speed equation on L157, the authors should greatly expand the magnitude of this sixth source of error/uncertainty, maybe by an order of magnitude.

We followed Varon et al. (2021) who prescribe a linear effective wind calibration for Sentinel-2 specifically, and thus did not try the log(U10) calibration.

New text: line 193 – 195	[…] and perform an effective wind speed calibration that only includes the pixels located above the source area in the plume mask. **Following Varon et al. (2021), we perform a linear regression of Ueff against U10m that is more appropriate for Sentinel-2-like instruments than the logarithm-based regression first proposed in Varon et al. (2018).** We obtain the following NS2-custom effective wind speed calibration with an outlier-resilient […]

In the revised manuscript, we expanded on the discussion of why the effective wind speed should exceed U10 in this specific case. These additions explain why in ideal conditions over land the effective wind speed slope is below 1, and why it is above 1 for the NS2 case

New text: line 176 – 179	Varon et al. (2021) provide an effective wind speed calibration model for Sentinel-2-like Earth imagers: Ueff = 0.33 × U10m + 0.45. **This IME effective wind speed calibration slope which is lower than 1 reflects the fact that the plume extent L, defined as the square root of the plume area, is smaller than the actual plume length for long narrow plumes observed over land. This definition of L is chosen for its simplicity and because the plume mask is ventilated by turbulent diffusion rather than uniform transport (Varon et al., 2018).** Besides, using this effective wind speed calibration implicitly assumes that the plume is observed in the same conditions as those used for the LES calibration, including for instance that the full extent of the plume is visible as per the given instrument sensitivity.

New text: line 197 – 201	This 1.88 calibration **factor is significantly different from the slope value given in Sect. 2.3.1, applicable for ideal conditions over land. Its value higher than 1 reflects a different plume definition compared to ideal conditions over land, and must be interpreted as methane excess observed above the area source under-representing the actual emission rate of the full area source. Indeed, only the downwind plume integrates emissions from the all the area source, not the concentration field right above it.** Actually, this IME effective wind speed calibration slope close to 2 is consistent with expectations from mass balance of a uniformly ventilated area source (wind direction above it is unique and not changing, a fair assumption at NS2 leak scale) as shown by Buchwitz et al. (2017).

Besides, we now provide in the Supplements the scatter plot of LES sampling points and Huber linear fit that were performed to obtain this effective wind speed calibration (Figure 13). It also gives the distribution of effective wind speed – fit mismatches which shows a 1.1 m/s standard deviation. This represents 12% and 10% of the averaged effective wind speed we obtain from the values sampled from ERA5, GEOS-FP, GFS and that we get from Bornholm airport for L8 and S-2B overpasses, respectively.

New text: line 231 – 233	**(6) We account for effective wind speed calibration errors by randomly sampling data – fit mismatch values from the distribution shown in the Supplements (Figure S8). By doing so, we implicitly follow the slightly non-Gaussian skewed distribution that these mismatches show.**

L173: What is the relevance of the uniformity of the ventilation?
We clarified this point in the revised manuscript.

New text: line 202 – 203	Actually, this IME effective wind speed calibration slope close to 2 is consistent with expectations from mass balance of a uniformly ventilated area source (**wind direction above it is unique and not changing, a fair assumption at the scale of the NS2 leak)** as shown by (Buchwitz et al., 2017).

L189: Why include four speeds and perturb each of them by 50%? The authors might actually be overestimating this source of uncertainty, since L193-L195 show that the wind speed range is not that large, especially on Sept. 30th.

We include four different wind speeds because this gives a better representation of the actual uncertainty in the wind compared to just perturbating one value. We adjusted the revised manuscript.

New text: line 220	(4) **Following Schuit et al. (2023),** we include four different 10-m wind speeds **to better account for wind speed uncertainty.** Three come from meteorological re-analysis products: the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 (Hersbach et al., 2020), the Global Forcasting System (GFS) from NOAA National Centers for Environmental Prediction (NCEP, 2000) and the Goddard Earth Observing System-Forward Processing (GESO-FP, Molod et al., 2012). Furthermore, we include the in-situ wind speed measured at Bornholm airport, which is located about 50 km away from the NS2 leak (IEM, 2023). For September 29th, we obtain wind speeds of 4.1, 6.6, 4.8 and 3.6 m/s from ERA5, GFS, GEOS-FP and airport measurements, respectively; and for September 30th, we obtain wind speeds of 5.0, 6.3, 6.3 and 5.7 m/s, respectively. **We randomly pick one of these four wind speeds.**

This 50% wind speed uncertainty was a conservative assumption, following Schuit et al. (2023). To better evaluate the actual uncertainty, we now compare ERA5, GEOS-FP and GFS products with wind speed measured bi-hourly at Bornholm airport for 2022. Over the three reanalysis products, we obtain an averaged standard deviation for "re-analysis – Bornholm airport" differences of 1.6 m/s. This amounts to 34% and 28% of the averaged 10m wind speed obtained from all these data sources at the time of L8 and S-2B overpasses, respectively. Overall, the first-order sensitivity index calculation that we provide show that this wind speed uncertainty contributes little to the emission result variance compared to the contribution of the empirical calibration of sea foam albedo spectral dependence uncertainty. We adjusted the revised manuscript.

New text: line 228 – 230	**(5) To account for wind speed error, we evaluate the differences between the three re-analysis models (ERA5, GEOS-FP, GFS) and in-situ measurements made at Bornholm airport for 2022. On average, we find a standard deviation of 1.6 m/s. We therefore sample the wind speed error from a Gaussian distribution with a 1.6 m/s standard deviation and centred on zero.**

L198: I cannot reproduce the numbers.
In the original version, the ensemble description had one slight imprecision: it lacked to specify that the perturbation on effective wind speed coefficients was applied at the same time on both coefficients, thus possibly causing trouble when trying to reproduce the numbers. In the revised manuscript, with the Monte Carlo ensemble approach, we randomly draw 1000000 members for each satellite, thus fixing this ambiguity.

L218: Re: "1M", is this 1 million? If so, please avoid the shorthand and I don't see why the random draws did not come from ~5 million members, but it might not matters. I simply need to know how the one million members were selected by the authors in order to assess whether this is a biased sample.

"1M" stood for 1 million indeed in the original version. We removed this ambiguous shorthand from the revised manuscript. Because we did not have the same number of ensemble members for L8 and S-2B (because of different minimum albedo interval lengths), we randomly drew (with replacement) 1000000 members from their respective ensembles to get the statistics of the L8 and S-2B average. Because results would change very slightly over different draws of 1000000 members, we reported the mean statistics of L8 and S-2B averages over 100 draws of 1000000 members.

We agree this was a cumbersome way to proceed. It is fixed in the revised manuscript, with the Monte Carlo ensemble approach, we now evaluate the statistics of L8 and S-2B average by combining our 1000000-member single satellite ensembles.

L236 (and in the abstract): hypotheses-> assumptions
We corrected this choice of word in the revised manuscript.

New text: line 17	This very specific NS2 observation case challenges some of MBSP and IME implicit **assumptions**, and thus calls for customized calibrations:

New text: line 63	This work first aims to show how Landsat 8 and Sentinel-2B observations of the Nord Stream 2 leak challenge implicit **assumptions** in methods usually applied for Earth-imager methane plume analysis and emission rate quantification.
New text: line 276	We have shown how the unusual observations of a sea foam patch surrounded by dark still sea (and clouds for L8) challenge implicit underlying **assumptions** in both the Multi-Band Single-Pass (MBSP) and Integrated Mass Enhancement (IME) methods.
New text: line 287	Overall, we see our work as a methodological cautionary tale illustrating how implicit method **assumptions** need to be considered and compensated for in unusual observation cases such as this one.

References

Varon, D. J., Jervis, D., McKeever, J., Spence, I., Gains, D., and Jacob, D. J.: High-frequency monitoring of anomalous methane point sources with multispectral Sentinel-2 satellite observations, Atmospheric Measurement Techniques, 14, 2771–2785, https://doi.org/10.5194/amt-14-2771-2021, 2021

Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, Atmospheric Measurement Techniques, 11, 5673–5686, https://doi.org/10.5194/amt-11-5673-2018, 2018.

Jia, M., Li, F., Zhang, Y., Wu, M., Li, Y., Feng, S., Wang, H., Chen, H., Ju, W., Lin, J., Cai, J., Zhang, Y., and Jiang, F.: The Nord Stream pipeline gas leaks released approximately 220,000 tonnes of methane into the atmosphere, Environmental Science and Ecotechnology, 12, 100 210, https://doi.org/https://doi.org/10.1016/j.ese.2022.100210, 2022.

Shine, K. P., Campargue, A., Mondelain, D., McPheat, R. A., Ptashnik, I. V., & Weidmann, D. (2016). The water vapour continuum in near-infrared windows–Current understanding and prospects for its inclusion in spectroscopic databases. Journal of Molecular Spectroscopy, 327, 193-208.

Zhang, Hankui K., et al. "Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences." Remote sensing of environment 215 (2018): 482-494.

Buchwitz, M., Schneising, O., Reuter, M., Heymann, J., Krautwurst, S., Bovensmann, H., Burrows, J. P., Boesch, H., Parker, R. J., Somkuti, P., Detmers, R. G., Hasekamp, O. P., Aben, I., Butz, A., Frankenberg, C., and Turner, A. J.: Satellite-derived methane hotspot emission estimates using a fast data-driven method, Atmos. Chem. Phys., 17, 5751–5774, https://doi.org/10.5194/acp-17-5751-2017, 2017.

Schuit, B. J., Maasakkers, J. D., Bijl, P., Mahapatra, G., van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D. J., McKeever, J., Jervis, D., Girard, M., Irakulis-Loitxate, I., Gorroño, J., Guanter, L., Cusworth, D. H., and Aben, I.: Automated detection and monitoring of methane super-emitters using satellite data, Atmos. Chem. Phys., 23, 9071–9098, https://doi.org/10.5194/acp-23-9071-2023, 2023.