Review on Pallandt et al., egusphere-2025-604

General comments

The study examines possibility of current and future observations to detect potential CH4 emission increase from Yedoma permafrost in the Arctic. It addresses important issue of whether we can detect any changes in Arctic CH4 emissions in future. Permafrost in the Arctic is in thread of thaw due to fast global warming and Arctic amplifications, from which a vast amount of carbon could be released. To examine this point, the study used GEOS Earth System Model based Observing System Simulation Experiments (OSSE) with scenarios with varying emission enhancement factors (i.e. increase in emissions from original values) and assimilating data from ground-based stations, TROPOMI and MERLIN satellites, and their synthetic data. The authors have chosen 2010 as baseline study year, which means that some ground-based stations have actual observed data, but not all. In addition, TROPOMI and MERLIN experiments are purely based on synthetic data and their generated uncertainties. The results showed that it is challenging to detect potential changes in CH4 emissions from Yedoma with current and future planned observation networks and current accuracy/precision of the GEOS model. This is worrying and raises challenges we must address urgently. I therefore recommend the paper has potential to be published, but would like to address a few points for revision.

- Considering that the scope of AMT, I would like to see more details of the specifications of the measurement systems, especially regarding differences between TROPOMI and MERLIN. It is unclear from current manuscript that why some differences arise, and how much of them are due to design of these instruments. Do they measure same quantity? Why uncertainties are difference in these two? For example, you mention that “snow and ice negatively affects the precision of both instruments, especially MERLIN” (L277) – why does this affect MERLIN more than TROPOMI? Why TROPOMI has no reliable way of sampling over open sea? Note also that TROPOMI is name of an instrument while MERLIN is name of a mission. Although it is readable in the current form, please revise and use the terms properly in the texts.

- As shown, the results of the study depends very much on generated uncertainty estimates (both of observations and transport model). Although authors try their best to guess the uncertainties, the results has to be interpreted with much care. Therefore:
  - the authors must make it clear that the study does not tell exactly the signal detection limit of the measurements/satellite retrievals, but rather of a transport model and on regional scale.
  - I would like to see more discussion on how the assumed uncertainties are comparable to original data (for ground-based stations and TROPOMI), other studies regarding the transport model (e.g. GEOS vs other models) and total uncertainties (e.g. those used in atmospheric inversion studies) – with focus on the Arctic regions.
  - Why did you test different transport model uncertainty for the MERLIN case, but not for the TROPOMI case? Do I understand it correctly that you did not add any transport model uncertainties for the TROPOMI case? If possible, I would like to see additional simulations with transport model uncertainties added to the TROPOMI case. In addition, please present results from the MERLIN case with different transport model uncertainties – I see that you compare between nature run and high uncertainty scenarios, but not between the two uncertainties… or did I miss something?
  - Please also mention what improvements are needed to reduce these uncertainties, especially regarding retrieval methods and transport models in e.g. Section 4.3.

Specific comments

Section 2.2: Could you add a table with a list of sites with site information (name, location), sampling method, sampling height, sampling time (if this varies per site; see my comments below), measurement error and total error (but if a uniform transport model was assigned to all sites, then total error would not need)?

L169-172: This is a bit unclear to me. You said that your database is limited to local time afternoon, but on the other hand, it is said that sampling was done during 9:00 to 19:00 LT. Could you clarify this point? Does the sampling time vary for each site?

L175-181:

- Here are three different error terms: "measurement error", "gaussian random error", "transport modelling error". However, without descriptions of them earlier, it is not easy to understand those terms and their differences. Please add descriptions.
- It is not clear what is meant by "This gaussian random error was scaled with a 95th percentile (P95) at 2 ppb (μ 0 ppb, σ 1.02 ppb)". You said "measurement error" was 2 ppb, but then the sentence explains "**this** gaussian random error". Do you mean e.g. "In this scenario, the gaussian random error was…"?
- I am confused with terms "95th percentile (P_{95}) at 2 ppb (μ 0 ppb, σ 1.02 ppb)" and "30 ppb (μ 0 ppb, σ 44.5 ppb)". Do you mean that you assigned random errors drown from normal distribution with mean 2/30 ppb and standard deviation of 1.02/44.5 ppb? Could you use more proper notations/terms? What does "95th percentile (P_{95})" do…?
- For TT_f scenario, do you mean that total observational error (measurement + transport model errors) was drawn from the normal distribution with mean 30 ppb and std 44.5 ppb, or is this just the transport model error?
- For how was the standard deviations (σ) set, i.e. why did you chose the values 1.02 and 44.5?

L194-195: What assessment did you do? Do you mean that albedo as a dominant factor for explaining variability in retrieval uncertainty?

L228: "We filtered out all fully clouded soundings"
Did you use same cloud screening as the one used to generate the TROPOMI soundings?

Regarding Sections 2.3 and 2.4:

- Please also describe how you have generated CH4/XCH4 values.
- You have described how the random errors are generated, but do you also add some systematic bias on top of it, or do you assume the systematic biases to be zero?
- What did you do regarding averaging kernels and pressure weighting functions for the gridded datasets?
- Please consider unifying the error terms used, and synthesise with terms in Section 2.2. I guess with precision, you mean "measurement error". Is there specific reason why the different term is used here?
- Could you consider adding plots showing spatial distribution of generated concentrations and uncertainties?
- Did you add any transport modelling error to the TROPOMI generated dataset? If not why?

Section 2.4: You have named two error schemes with abbreviation for the surface inversions, but not for MERLIN. Why? Just to improve readability, it could be useful to add such and use it in the main text similarly to TT_i / TT_f notations.

L242:"Here a, b and c are constants which were set to 20, 0.2 and 70 respectively to match Fig. 2 of Bousquet et al. (2018);"
Sorry, I do not quite understand why you chose those values, and what each variable means, even after looking at Fig. 2 of Bousquet et al. (2018), which shows maps with varying values globally. Could you add short descriptions of the variables and a bit more detail why you chose those values?

Eq. 3: Where does "142" come from, and why do you need this scalar?

In Eq. 3 and L242, does "*cfrac*" mean same as "*cf*" in section 2.3? If so, please use the same term.

L257: How did you come up with the lowest Fe of 1.06? Why you chose 80 steps in between? Could the steps be smaller or larger?

L258: Do you examine only after 1st March when you start increasing the emissions, i.e. first bin with size 7-day is 1st – 7th March?

L289-292: Is this the case without transport model uncertainty? Throughout the result section, please make it clear which scenario you are talking about – not only of the datasets, but of uncertainties.

L319-321: I do not quite understand. Are you arguing that with smaller sample sizes, the detection limit is lower in general? But then it does contradict with the findings that detection limit is smallest in the *Full* dataset. In case of ground-based stations, I kind of understand, as it is the condition that "at least one site detect changes". But in the satellite case, the dataset is spread over the region, and there are data not close to Yamada as well. I wonder how much the number of samples in general can explain the differences in the detection limits.

- Why smaller sample size necessary lead to lower detection limit?
- Number of soundings over ice is much larger than those over land in March-April.

Section 3: Earlier, you mentioned that "we focus on the 112 day bin" (L278), but it seems that sections 3.3 focus on 28-day bin. How about Section 3.2 which bin size results are you talking about?

Comments to Figures and Tables
Figure 2:

- Please add also scenario when lowest Fe was used.
- Either as a subpanel or another figure, please add time series of baseline Yedoma wetland emissions and total emissions in a study domain. Please also consider adding generated XCH_4 values.

Figures 3, 6, 7, 8:

- "Non linear y-axis of all 80 Fe steps." I think this could be slightly misleading as y-axis ticks do not show all 80 steps. Perhaps rewrite simply as e.g. "Non linear y-axis shows flux enhancement (Fe)"?

Figure 5: Does this show number of original data or those processed and used in OSSE?

Fig. 6-8: Why x-axis starts in January when emissions should be the same in natural and enhanced scenarios? This is also related to my question regarding L258.

Y-axis labels in Fig. 3 & 6-8: Could you consider modifying them as e.g. "Flux enhancement factor (Fe)"? I think it would be more informative this way for busy readers.

Figure 7 and 8 captions: “the q-value of the comparison between the baseline and enhancement scenario”.
Could you consider modifying the phrase as e.g. “the statistical significance (q-value) of the differences in the detection limits between the baseline and enhancement scenarios”?

Table 1:

- This shows statistics/values not of the original data, but generated values. Please indicate it clearly in the caption.
- What is “mean random error” exactly? Does it contain also transport model error, or do you mean measurement error (i.e. “precision” in Section 2.3 and 2.4)? Please unify terms with the main texts to avoid confusion.
- Could you also consider adding mean XCH_4 mole fractions, and not only the errors?
- I recommend you remove the last two rows from this table, and make another separate table regarding transport model errors added to the surface, TROPOMI and MERLIN data.

Technical comments
L228: “Total column soundings were performed” → “Total column soundings were generated”

L367: “not unreasonable” → “reasonable”

L412: “transport errors” → “transport modelling errors”