We would like to thank both reviewers for taking the time to review this manuscript and their helpful and fair comments. As suggested, we reran the model based on updated TROMPOMI products, updated most figures and added a few new ones, and in general expanded the text both to clarify the method and choices made as well as put it and the nature of the OSSE in context. This greatly improved the manuscript and we hope the reviewers agree. Below we answer (in purple) the questions and comments of the reviewers (in green).

Kind regards,
Martijn

Reviewer 1

General comments

The paper uses outdated versions of the EDGAR emission database and TROPOMI retrieval products (see specific comments), without a proper justification. At the very least, such a justification should be added to the paper, but the paper would be more relevant if the OSSEs were based on recent datasets.
This is fair, in the specific comments we justify the use of an older EDGAR version and we redid the entire TROPOMI analysis using V2.0 of the WFMD retrieval.

Specific comments

*) page 3, line 125: it is mentioned that the model as a 0.5 degree horizontal resolution. But the horizontal extent is not mentioned, is it global or regional? Figure 1 and 5 only show latitude >~ 50N.
The model ran globally, for the analysis we only considered north of 50° North. We updated the text to reflect this.

*) page 3, line 128: Why use EDGAR v4.3.2 (which is apparently from 2017, even though the cited publication is from 2019. See https://edgar.jrc.ec.europa.eu/archived_datasets) if more recent versions of the database are available (such as the 2024 version EDGAR_2024_GHG)?

We thank the reviewer for this comment. In our case, the GEOS runs were carried out back in 2020, when EDGAR v4.3.2 GHG was the latest available version. These runs used the GEOS model configuration that is reported in Sweeney et al. 2022, ACP (although that paper itself was submitted back in 2020!). In the Sweeney et al. paper, the GEOS model simulations were extensively validated against aircraft observations over the northern high-latitudes. Since such a robust evaluation was already conducted with a specific GEOS model configuration, we decided that it would be prudent to stick to that setup for our study. We agree with the reviewer that since then more recent versions of EDGAR data are available but if we were to switch to a new EDGAR data version, we would have to go back and redo all the comparisons against independent data, make sure the model configuration and setup is robust – all of which could further delay getting this important study out to the community. Also, the GEOS model runs themselves take significant time and computing resources to run, hence we decided to stick to the previous EDGAR version rather than update it to the latest version. We would like to point out though that we do not expect our overall findings and conclusions to change since the EDGAR data provides information about agriculture and waste, fossil fuels and biofuel sectors, which are important for the total CH_4 signal, but not necessarily for the signal from the Yedoma region.

*) page 3, line 136: Why pick 2010 as baseline year? Later TROPOMI data from 2019 is used for the synthetic data. So you could have used actual TROPOMI data with measured cloud fractions etc..

When this was study was initiated the actual TROPOMI data were not yet available. The year 2010 was chosen to align with other GEOS runs. And since all generated data was to be synthetic, the exact year was of minor importance.

What would change if you modelled a longer time period?
In case of a burst of only one season we saw in one test run that ran for an additional year that atmospheric mixing will greatly reduce our ability in the following year without additional fluxes to detect differences. With continued enhancement we expect the lower detection thresholds to improve, though it might not necessarily improve our ability to pinpoint sources.
*) page 3, line 137: do you only amplify the Yedoma gridcells and keep the emission from the non-Yedoma grid cells constant?
Only wetland fluxes in Yedoma flagged grids are enhanced. We now clarified in the text that other fluxes are unaltered.

*) page 5, line 187: although the version numbering of TROPOMI SRON and operational retrievals could be improved, I think that v0017 is an older version of the SRON retrieval product. Lorente et al. (2022, https://doi.org/10.5194/amt-2022-255) describe important updates to the algorithm improving the retrievals. These updates have since then been incorporated into the operational (reprocessed) product. The question is therefore, why use an older version of the product?
*) page 5, line 189: Why use version 1.5 of the WFMD product? Version 1.8 has been available for years (e.g. https://doi.org/10.5194/amt-16-669-2023) while the most recent version is 2.0: https://www.iup.uni-bremen.de/carbon_ghg/products/tropomi_wfmd/index.php
Frankly, the original analysis happened so long ago that the mentioned products were up to data at the time. Though as suggested we revaluated the products and opted to redo the entire TROMPOMI analysis with the V2.0 of the WFMD product.

*) page 5, line 189/190: The sentence "Both of these products only contain successful retrievals." is incorrect. These products do contain e.g. non converged retrievals, but you can (and probably should) select only the successful retrievals with the provided quality flags.
Indeed, this was an error. We updated the text with the actual flags that were used.

*) page 5, line 199: "... fitting a curve to reported uncertainties..." Please provide a plot with this fit.
This plot is now added to the supplement (Fig. S1) and referenced in the text

*) page 5~7, sections 2.3 and 2.4: How is the model sampled for satellite observations? Is the averaging kernel taken into account?
We now describe in the text that Pressure-weighted column averaging was applied to generate these modelled samples, which is comparable to the actual TROPOMI averaging kernels and a nominal MERLIN weighting function. Refer to the supplement for more details, including added Fig. S3.

*) page 7, line 256-257: wrt. the t-test, is the test statistic used in a one or two-tailed significance test, and what about the null and alternative hypotheses?
Indeed, a few more details are required:
We compare the nature run with the Yedoma thaw scenario for each of the seven sampling and error characterizations listed above using an array of two-tailed t-tests to detect any difference with as alternative hypothesis that no detectable differences are present. We opted for two-tailed t-tests since in reality we would not know if at a certain point or time a flux would increase or decrease.
*) page 11~13, figures 6~8: why use 28-day bin sizes while in lines 277-279 (page 7) it is mentioned that "for the remaining evaluation we focus on the 112 day bin..."?

Indeed, this is somewhat confusing. These 28-day figures have a good balance between temporal resolution while still showing fairly optimal detection limits. We updated the text to indicate we do not exclusively asses the 112-day bins.

Technical corrections
Accepted all

*) page 1, line 3: affiliation of last author (Gockede) is missing.
*) page 1, line 15: please change satellites to satellite instruments.
*) page 2, line 85: since this is the first mention after the abstract of MERLIN, please provide the full instrument name in addition to the acronym.
*) page 6, table 1: fix the vertical alignent of the cell "Ice, TROPOMI" (containing the number 141461)