Review 2

The replies to the comments are highlighted in green.

The authors present a study that aims at assessing what enhancement level of Arctic CH4 emissions may be reliably detected, and spatially attributed, based on GHG mixing ratio observations from the pan-Arctic tall tower network. Their approach uses atmospheric transport modeling with FLEXPART to first general synthetic time series of mixing ratios that reflect the changes in the atmosphere following surface flux enhancements at selected regions. In subsequent steps, these synthetic time series are then used as input for atmospheric inversions in an attempt to quantify, and spatially attribute, the surface flux rates of CH4 corresponding to the chosen emission scenarios. Since in this synthetic setup the 'truth' is known, this approach allows to quantify how well the inversion-based posterior fluxes agree with the true emissions, at what level of flux enhancement the higher fluxes are significantly different from the baseline, and how well the flux trends are assigned to the correct target regions. Based on these metrics, the authors conclude that substantial flux enhancements are required for a reliable detection, particularly in regions with sparse observations, and that a mis-attribution of the flux signals is quite common.

The overall objectives that Wittig et al. aim at are highly relevant – even in the absence of a 'methane bomb' scenario, enhanced GHG emission rates from degrading Arctic permafrost can be expected under future climate change, and a monitoring system that would reliably pick such changes would therefore be very useful. The in-situ atmospheric GHG monitoring tower network, in combination with atmospheric inverse modeling, is a suitable tool for this purpose, but due to the sparse network coverage the sensitivity of this tool towards future changes is still uncertain. The approach used within the context of this study, i.e. generating synthetic datasets with a known truth that allows to assess how well trends are quantified, and how reliable the spatial attribution of fluxes is, is well suited for this purpose. Unfortunately, some settings in the inversion setup seem to be over-simplified, so that even though the qualitative results may be solid, most quantifications are very questionable.

I see 3 major issues that compromise the findings presented in this work:

First, the authors produce synthetic mixing ratio observations as a prerequisite for conducting inversions for future emission scenarios, but they do not apply uncertainties when using this information in the actual inversions. Or rather, such uncertainties are not described in the presented paper, but based on the statement on p.11, ll.17-18 (These figures reflect an ideal case where uncertainties in the inversion system are minimized) I presume that none were applied. If this is the case, then the same transfer functions were applied in forward (to produce the synthetic data) and in backward modes (to execute the inversion). This is an over-simplification of the situation in a regular atmospheric inversion, where model-data mismatch uncertainties such as e.g. transport, mixing, or aggregation errors are a key component that make the links between surface processes and atmospheric observations much more challenging. Accordingly, all quantitative findings presented herein, such as years until detection of a change, or emission thresholds until detection, are highly questionable, and detection limits are likely low-biased.

Indeed, we do not apply uncertainties on the synthetically generated observation (in forward mode). We are aware of this aspect and chose this method purposefully. The aim of our work is to evaluate the current stationary observation network in the Arctic region with regard to the detection of a possible methane bomb. Our main conclusion is that the network is not fully adequate for this purpose and is limited in its ability to detect small changes in emissions in the Arctic. Introducing uncertainties on the observations would lead to a similar conclusion, but with an even higher detection threshold.

We propose to clarify in Section 2:

“Theoretically, the synthetic observations ***y*** should be perturbed by an error $\in$ (with a Gaussian distribution, following the matrix **R**), accounting for measurement errors, as well as other uncertainties such as transport and aggregation (described e.g. by Szénási et al., 2021). In our approach, we deliberately disregard these errors in order to obtain optimistic results and assimilate optimal measurements to analyse the best possible detection of different observation networks (Section 3.2) regarding a methane bomb event.”

Additionally, we propose to add a paragraph summarising our general approach including clarification on the synthetic observations:

“In order to implement this work, we apply hypothetical trend scenarios on different CH_4 emission sources to simulate a methane bomb in different regions located in high northern latitudes. By combining these emission scenarios with the extrapolated output of an atmospheric transport model, we obtain synthetic CH_4 mixing ratios for the current observation network in the Arctic and Sub-Arctic as well as for an observation network extended by possible additional sites. These synthetic observations subsequently serve as input data for the inverse modelling setup in order to identify a temporal threshold of possible detection and to analyse regional differences in the ability of the two networks to adequately detect and localise increasing CH_4 emissions. Since we assume optimum quality and availability of the measurement data, the results obtained represent a best-case scenario for the detection of an Arctic methane bomb using exclusively in situ observations.”

As well as adding to the conclusion:

“In this approach, we have made the optimistic assumption of excellent quality and availability of measurement data. The results presented therefore represent the best possible scenario for detecting a future Arctic methane bomb.”

Second, the fact that pan-Arctic posterior fluxes are systematically low-biased, compared to the ‘true’ fluxes in this synthetic experiment, suggest that the chosen setup is compromised. The best explanation I could come up with to interpret this phenomenon is that the correlation length scale chosen for FLEXPART does not allow to reproduce the steep gradients in regional flux patterns that emerge when emissions only in one region are ramped up extremely, while the neighboring regions stay at their low prior values. Such gradients obviously pose a major challenge to any inversion framework fed by sparse observations, where gaps in between monitoring sites need to be interpolated based on assumed spatial relationships within flux fields, necessarily producing (to a varying degree) smooth result surfaces. There may be other factors at play here, but emission peaks within target regions that are systematically underestimate, while adjacent regions have a high bias

in fluxes, may be related to this. In any case, the problem requires further investigation, and in-depth discussion, both of which is currently lacking in this manuscript.

The transport model FLEXPART used in this work does not contain a modifiable function regarding the correlation length. Here, we use the footprints obtained from simulated backward trajectories to determine both the synthetic observations as well as their modelled equivalents, which serve as input for the inversion framework. We would assume that your comment about the correlation length refers to the spatial correlation of the prior error covariance matrix in the inverse modelling set-up, which is 500 km (mentioned in Figure 1). Information on the spatial and temporal correlations have been added in Section 2:

“The off-diagonal elements of the prior error covariance matrix are thereby determined by applying spatial and temporal correlations of 500 km and 7 days, respectively”

For a more comprehensive description we refer to our previous work (Wittig et al, 2023) to avoid unnecessary repetition.

Furthermore, as described in Sections 3.5, we did not only attribute a steep increase in methane emissions in each of the individual sub-regions separately, we also increased the emissions in much larger regions (see supplements, page 4, Figure S1). As demonstrated in Section 4.1, similar discrepancies between the obtained posterior results and the assumed *true* state of the fluxes were obtained.

In addition to that, our inverse modelling set-up also optimises the background concentrations alongside CH_4 fluxes. This has briefly been mentioned in the description of the state vector in Section 2. Since the prior background estimations are not perfect, part of the missing CH_4 mass may be compensated by increasing the posterior background concentrations.

We propose to add the following explanation in Section 2:

“In our analysis of the detectability of elevated Arctic CH_4 emissions (Section 4), we examine how accurately the truth is captured in the posterior emissions of different regions and whether these elevated fluxes are localised in the right area. By design, our inverse modelling system will try to fit additional fluxes by adding CH_4 emissions in the Arctic region, but possibly not at the correct location. Since, as described above, the background mixing ratios are also included in the control vector $\mathbf{x}^b$ and consequently optimised in the posterior state, part of the missing CH_4 mass is likely to be compensated by increasing the background, hence generating a low-bias in the posterior emissions.”

We have additionally highlighted the optimization of the background in Sections 2:

“Here, $\mathbf{x}^b$ also contains information on the initial CH_4 background mixing ratios (described in Section 3.4), which are therefore optimised in addition to the CH_4 fluxes.”

And Section 3.4:

"However, since an exact estimate of the background mixing ratios remains challenging and the calculated background concentrations do not provide perfect estimates, the background mixing ratios are optimised together with the CH_4 fluxes (see Section 2)."

Third, I find several issues in the statistical measures used to evaluate flux trends:

- Equation (4) compares prior fluxes (in 2020) to posterior fluxes (in future years) and relates them to observational uncertainties to assign a detection limit. This calculation is only valid if prior and posterior fluxes in 2020 are exactly the same, for every target region. Since this study is based on synthetic data, there should not be a major adjustment between prior and posterior in the absence of a trend in emissions; however, since the observational network is sparse, there is reason to assume that prior and posterior are not identical even in 2020, which may lead to systematic difference even when aggregating fluxes by region. Without demonstrating that priors and posteriors are identical on a regional basis, the posterior fluxes in 2020 should be used as a reference (emis_b_2020), instead of the priors.

  The prior fluxes in 2020 are indeed identical to the truth. The first year in which the true emissions deviate from the prior emissions is therefore 2021. This has been shown in equation 3 (Section 2). To further clarify this fact, we propose to add in Section 2:

  "This trend was only applied from the second year of the study period (2021), in the year 2020 the truth is identical to the prior state."

- Equation (5) compares 'true' to 'posterior' fluxes to quantify how much of the emission enhancement is actually captured by the inverse model. This is called 'detected trend magnitudes' in the section header, so the intention is obviously to quantify how much of the trend is captured. However, equation (5) compares absolute fluxes, not flux enhancements since 2020. For quantifying how much of the trend is actually captured by the inversion, I would find it more convincing to calculate how much the 'true' flux changed since 2020, and how much of a change is seen between the posterior fluxes over the same time span.

  In this analysis, we want to determine how well the *true* trend (which is known in our case since we determined it) is detected in each region by the inversion, not which total trend each region shows since 2020, as this would not give us any information about the performance of the observation network. Therefore, the deviation of the posterior state in the threshold year from the truth is a good indicator, since, if the trend would be adequately detected, the difference between the posterior emissions and the truth would be minor.

  For clarification, we propose to state at the beginning of 4.2.2:

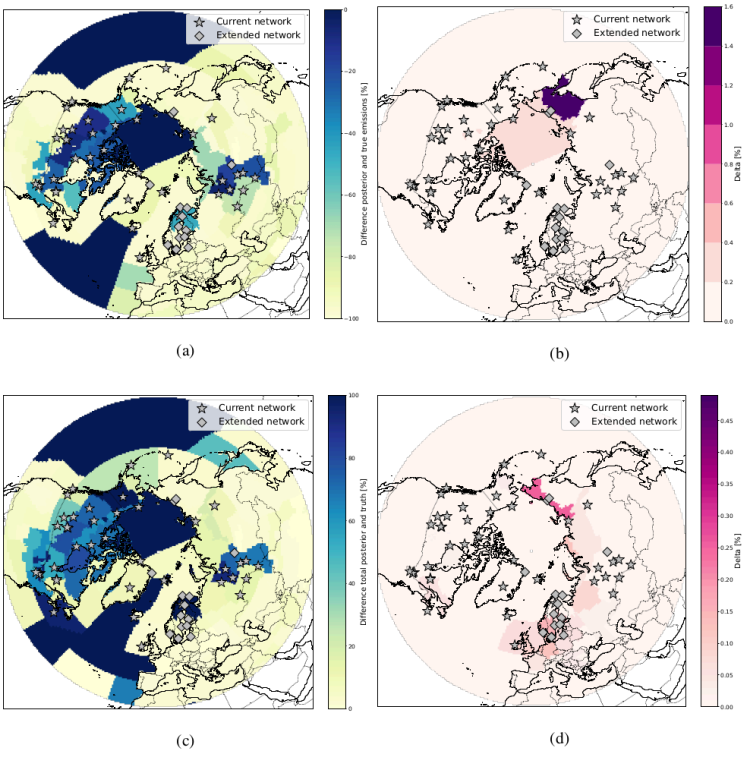
“Subsequently, we want to examine how well the previously determined trends of 20% increase in wetland emissions and 100% increase in oceanic CH_4 emissions, respectively, are captured in each of the corresponding sub-regions.”

In addition to that, we propose to extend our analysis to determine how much of the increase in CH_4 fluxes is detected over the whole Arctic domain by the inversion when increasing the CH_4 emissions in one of the sub-regions (see following comment). We therefore extended Figure 7 by two subfigures (now Figure 6c and 6d in the edited manuscript) and added the following description:

“Additionally, in order to determine the share of the truth detected by the inversion, we calculate the detection ratio $K_{j,r}$. Hereby, the posterior increment in all regions $\Sigma\Delta emisaj,r$ in the threshold year j is divided by the the true increment $\Delta emis^{t}_{j,r}$ in region r:

$$K_{j,r} = \sum \Delta emis^{a}_{j,r} / \Delta emis^{t}_{j,r}$$

with $j \in [2021, 2055]$ and $r \in [1, 121]$. Hence, we analyse how much of the true increase is detected, independent from the location it is attributed to, when increasing the CH_4 emissions in one of the sub-regions. Higher values indicate that a larger share of the true emissions is detected in the posterior emission, distributed over the whole pan-Arctic domain. Figure 6c shows that the detection ratio is generally higher when the true emissions are increased in regions with a dense observation network (such as North America), with values of up to 100 %. Similar to the relative difference (Figure 6a), the high detection ratios in the oceanic regions are due to the absence of trends in the true emissions, since the CH_4 emissions in these regions are nearly zero [...] Regarding the comparison of the detection ratio of the two networks, shown in Figure 6d, the improvement is even smaller with a maximum of 0.3 %.”



- I cannot really follow the logic behind equation (6), even though the objective to quantify mis-attribution is highly relevant. Why is the 'delta_emis' measure used here? The authors state themselves in Section 4.2.2 that a good 'delta_emis' factor, i.e. with a value close to zero, indicates that the posterior is very close to the true emissions. Now when dividing the sum of 'delta_emis' in other regions by the 'delta_emis' from the study region, if the latter value is close to zero the result would be rather high increment ratios ..?? So why use the 'delta' measure in the first place here? I would find it much more intuitive if the authors first quantify how much integrated pan-Arctic flux budgets were increased when raising fluxes in a single region, i.e. how much of that 'true' signal is actually detected by the inversion, no matter where exactly. Next, you should simply quantify what fraction of that enhancement is attributed to the target region where fluxes were enhanced, and what fraction lies outside.

  The parameter $\Delta emis^{a}_{j,r}$ used in equation 6 is not equal to $\Delta emis_{j,r}$ calculated in equation 5 (hence the use of the exponent "a"). It is rather the increment of the posterior fluxes in the threshold year since the year 2020. We have edited the description of this parameter to avoid misunderstandings:

  "$\Delta emis^{a}_{j,r}$ and $\Delta emis^{a}_{j,i}$ hereby represent the difference between the posterior CH_4 emissions in the threshold year j and the true emissions in the year 2020 in the corresponding region r or i, respectively."

  Thus, we calculate the sum of the posterior increments in all other regions divided by the posterior increment of the targeted sub-region. This value should ideally be zero, since this would indicate that no posterior increment would be detected outside the region in which the fluxes were increased. Higher absolute values on the other hand indicate that the increment outside the target region is higher than in the region itself.

  Your suggestion to quantify the integrated pan-Arctic flux budget when raising fluxes in one of the sub-regions was implemented in Section 4.2.2 (see previous comment).

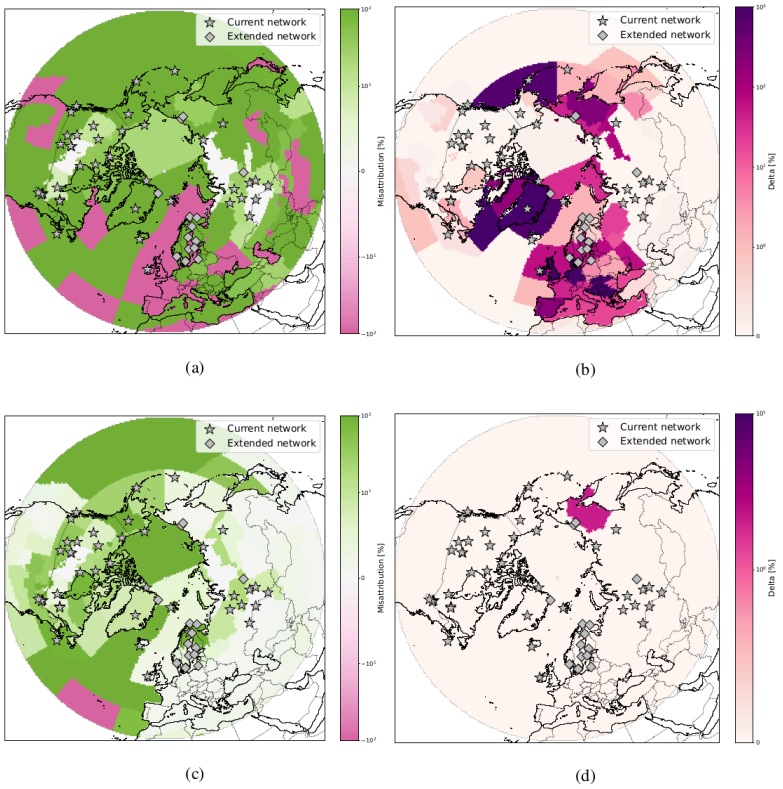
  Moreover, we propose to extent Figure 8 by two subfigures (now figures 7c and 7d in the edited manuscript) with the following analysis in order to determine the misattribution of the true fluxes:

  "In addition to the posterior increment ratio, we compute the true increment ratio $\kappa^{t}_{j,i}$ for each sub-region i:

  $\kappa^{t}_{j,i} = \sum \Delta emis^{a}_{j,r} / \Delta emis^{t}_{j,i}$

  for the threshold year $j \in [2021, 2055]$ and the region $r \in [1, 121]$ $r \neq i$. $\Delta emis^{t}_{j,i}$ is hereby defined as the difference between the true CH_4 fluxes in the threshold year j and the truth in 2020 in the corresponding region i. The closer the value of $\kappa^{t}_{j,i}$ of a specific region is to zero, the less true emissions are misattributed to other sub-regions. The true increment ratios are shown in Figure 7c. Similar to the posterior increment ratios, the fluxes are generally less misattributed when the true

emissions are increased in continental areas with available observation sites, especially in Siberia and Canada. The improvements from the extended observation network are smaller regarding the true increment ratio (see Figure 7d) in comparison to the posterior increment ratio, with only one region in eastern Siberia showing a clear improvement of around 10%."



Additional comments:

- I don't find the flow charts (Figs. 1, 4) too helpful in the current format.

  Figure 1 and 4 have been edited. Figure 1 has been included in the supplements instead and additional description has been added to Figure 4.

- Using emissions, or emission thresholds, in absolute numbers (e.g. Tg/yr) is misleading, since the size of the regions is variable, and unknown to the reader. Fluxes would be more intuitive if normalized by area.

  For our purposes, we consider it relevant to have absolute numbers for the emission threshold. Even though the sub-regions have varying sizes, we are more interested in the "burden" on the atmospheric CH_4 concentrations by a given region.

- When presenting the 'inversion method' as Section 2, some details are missing. Maybe it would be better to place this after the 'material' section.

  We would prefer to leave the methods before the material, as we repeatedly refer to the content of the inverse modelling section in the material section. We think it is difficult to understand why, for example, the synthetic observations are created

without knowing the background of the inverse modelling set-up. Additional references to the material sections have been added in the methods section for clarification.

- In Section 3.2, definitions are not 'clean', since the assumption of continuous data everywhere already upgrades the 'current' network. Also, some more details on the networks, e.g. total number of sites, or regional distribution, should be added to the text.

  We propose to clarify and add:

  "The term "current" refers hereby only to the location of the stations. This network, as used in this study, already provides additional data compared to the actual observations available from these sites. This is because, as stated before, we assume continuous measurements where currently only flask measurements are carried out. The current network contains hereby 40 stations in total, whereby the majority (26 sites) of the sites is located in North America (Canada, USA and Greenland). 10 observation sites are located in the Russian Arctic and Sub-Arctic and 4 sites in Northern and Western Europe (Finland, Norway, Ireland and Iceland). [...] The extended network expands the current network by 16 observation sites. The majority of these stations, 11, are located in Northern Europe (Sweden, Finland, Norway, Lithuania and East Russia), 3 in Central and Western Russia and one station each in Canada and Greenland."

- In Section 3.4, more information on the setup of FLEXPART and the optimization strategy would be helpful.

  The configurations used for FLEXPART have been described as detailed as possible in the current manuscript (page 7, lines 7 to 12). There are no other parameters in this transport model that are relevant for the inversion and that we could specify further. For more detailed information on the function of FLEXPART version 10.3, we refer to Pisso et al., 2019.

  Additional information on how the FLEXPART simulations have been used as input for the inverse modelling framework have been added in Section 3.4:

  "The so-called footprints obtained by sampling the near-surface residence time of the various backward trajectories of the virtual particles are subsequently used to determine the CH_4 mixing ratios per methane emission sector (Section 3.3) and sub-region (Section 3.1). The footprints define hereby the connection between the methane fluxes discretised in space and time and the change in concentrations at the observation site (Seibert and Frank, 2004). To obtain a time series of modelled CH_4 mixing ratios, a time series of footprints is integrated with discretised CH_4 flux estimates. As described in Section 2, in the inverse modelling framework, the modelled CH_4 mixing ratios obtained from the FLEXPART footprints are included in the observation operator **H**. In this study, this matrix is used for both the calculation of the synthetic future observations (shown in Equation 3) based on future emission

scenarios (see Section 3.5) as well as their modelled equivalents based on prior emission estimates.”

The optimization strategy is an analytical Bayesian inversion framework (equation 1 and 2) as described in Sections 2. This framework, including the components as well as the corresponding input data, have been extensively described in the manuscript and the references provided here cover the main points.