How to trace the origins of short-lived atmospheric species in the Arctic

Reply to reviews

We thank both the reviewers for their thorough readings and constructive and helpful comments. Our answers are given below in blue, while changes to the manuscript are given in italics. Unless otherwise stated, line numbers refer to the manuscript submitted before reviews.

Referee 1 (RC1)

The submitted paper, titled **"How to trace the origins of short-lived atmospheric species in the Arctic"**, investigates the origins of particles and trace gases in rapidly changing polar climates, with a focus on aerosol-cloud interactions. The authors highlight the limitations of current backtrajectory models like FLEXPART in identifying emission sources accurately, emphasizing the need for improvement due to the impact of aerosols on polar clouds and climate modeling.

To address this, the study combines backtrajectories from FLEXPART with tracer simulations from WRF-Chem, enabling a precise evaluation of source detection methods. They present a new approach based on backtrajectory analysis, to improve source identification accuracy through parameter sensitivity studies and validations using Arctic aerosol campaign data. Their results demonstrate the flaws of traditional backtrajectory analysis and the skills of the revised method in correctly identifying sources of methane sulfonic acid and black carbon.

The methodology presented in this paper appears robust and well-developed, addressing key challenges in tracing the origins of short-lived atmospheric species in polar regions. The combination of WRF-Chem tracer simulations and FLEXPART backtrajectory analysis represents a significant step forward in improving source identification accuracy. The results are convincingly validated with observational data, making this study a valuable contribution to the field of atmospheric sciences.

I recommend this paper for publication, subject to the authors addressing the minor comments outlined below.

We gratefully thank Referee 1 for helpful comments regarding our submitted manuscript. The comments helped us to question several points of our analysis and therefore to clarify the manuscript consequently.

Please find our answers below.

1) The choice of a 50 km × 50 km grid resolution for FPES calculations might limit the method's ability to resolve emissions from localized or highly dynamic sources such as ship traffic. Given the transient and narrow spatial footprint of such sources, the averaging approach inherent in the method could dilute the contribution of mobile emissions and introduce overlap with nearby stationary sources. Have you tested the sensitivity of your method with a 25x25km or smaller grid?

Reply: FLEXPART is a Lagrangian dispersion model, its trajectory tracking is not affected by the simulation domain resolution like it would have been with an Eulerian approach. However, once plotted on the 50×50 km² WRF grid, small signals would appear diluted in the grid cell. Their non-detection by the method would be due to the weakness of their emission rather than to the grid resolution.

This averaging approach makes the detection of individual ship tracks unlikely. It is not excluded that recurrent commercial shipping paths could be detected. In the reply to comment n°3, we suggest additional filtering to achieve that goal.

We did test the method on localized sources. But, although the detection signal included the emission point, it spread over large areas, making the precise identification of the source impossible. A way to achieve a more precise detection is to combine the analysis of several simultaneous measurement series of the emitted plumes sampled at geographically distant stations, as a way of triangulating the emission source. This can easily be done with satellite data, by interpolating time series of the studied species at different locations in the region of interest, but the limitation will be provided by the horizontal resolution of the satellite product. This method is described in Sect. 3.5.

2) The optimized cutting threshold on FPES is set to 2%, which the authors appropriately note in the discussion cannot be generalized to other regions or surface sources. To enhance the potential for generalizing these results, I suggest also to translate the FPES values into a quantifiable number of trajectories contributing to a given FPES value. This would provide a more universally interpretable metric for future applications. When stating that the purpose of filtering the FPES is to remove isolated backtrajectories, could the authors clarify what constitutes an 'isolated trajectory'? Does it refer to a single trajectory, or a minimal number of trajectories in a given grid cell?

Reply: Thank you for your relevant suggestion. Indeed, we tested various filtering thresholds on the number of trajectories passing through a simulation grid cell. Actually, both methods are equivalent since we set up a threshold that corresponds to a fraction of the total count (here 2%) of launched trajectories (or emitted tracers). In practice, both ways give the same results.
In order to make the method more general, the following sentence has been added to the section that introduces the filtering of FPES, Sect. 3.3, L255:
"*It is worth noting that applying a filter on FPES or PES values is on average equivalent to setting a threshold on the number of trajectories passing through a grid cell.*"
And the first two sentences of section 4.3 on sensitivity to FPES filtering have been changed to:
"*The improved ratio method (Sect. 3.3) incorporates a filter to exclude the grid cells with the lowest FPES values — i.e., those penetrated by the fewest trajectories. This serves to remove the least statistically significant results, as well as particle dispersion modeling imprecision near the domain boundaries.*"

We called "isolated backtrajectories plumes" (L256) the low values of FPES, which, as said above, is equivalent to a low number (under a given threshold, here corresponding to the 2% lowest FPES) of trajectories passing through a given grid cell. We replaced the expression by "FPES", since it is more accurate and clearer.

3) Could the authors elaborate on how their method would perform in isolating the impact of ship traffic, particularly in areas where shipping lanes are adjacent to other emission sources, such as coastal or industrial regions? Would finer grid resolutions or additional filtering parameters improve the reliability of source detection for such cases?"

Reply: The performed tests showed imprecision over coastal areas, as expressed about the overflows/shadowing of oceanic signal over continental regions in the MSA application example (Sect. 5.2). These overflows seem to be inherent to the method, as well as independent to the domain resolution since they are most of the time several grid cells wide. In the conclusion (lines 487-494), we attribute this limitation on detection precision to the intrinsic uncertainties of the backtrajectory modelling. A part of our evaluation work on source detection with backtrajectory approaches was to estimate this lack of precision. Figure 4b presents how much of overflow one can get for the three types of emission sources and for the five measurement stations. Unfortunately, the results show that a perfectly precise source detection can not be achieved.

In order to detect ship tracks, one could use the method with a filter that removes trajectories which have footprint PES over continental areas. In this way, the emissions from continental regions would be cleared off. The filtering on trajectory count, or on FPES, should be adjusted consequently. One limitation that may occur from this approach is that plumes from shipping emissions might be mixed within polluted plumes coming from the continent and advected over the shipping lanes before being sampled at a certain measurement station. The result of this filtering approach would be to generally attribute the source of this mixed pollution to ships only.

Past studies have investigated ship track emissions with Lagrangian modelling. In Marelle et al. (2016), the authors used FLEXPART-WRF in forward mode to quantify the impacts of ship traffic on air quality and radiative budget along the Norwegian coast. We think their approach constitutes a good way to investigate ship traffic emissions.

Referee 2 (RC2)

Summary

The manuscript 'How to trace the origins of short-lived atmospheric species in the Arctic' by Anderson Da Silva and co-workers analyses the robustness of a previously used, qualitative source attribution (localisation) method by applying it in what would be called an observation system simulation experiment (OSSE) in the inverse modelling community. They can show that the method suffers from several limitations, which they then try to remove by applying additional filtering (baseline subtraction, removal of areas with low residence time). Similar techniques have been used in previous studies, which is not reflected in the current manuscript and gives the impression that the 'improved' method is wholly novel. Nevertheless, one novelty of the 'improved' method is the combination of high concentration footprints with low concentration footprints. Furthermore, the application of the 'improved' method to real observations of two atmospheric compounds confirms the general suitability for the presented setup of Arctic measurement stations and the compounds treated (short-lived with exponential decay). Although, the manuscript is mostly clearly written it suffers from a lack of clarity on past and present model approaches for source attribution, both in terms of awareness of existing methods and description/application of transport simulations. Furthermore, the shortcomings of qualitative source attribution methods are well known in the greenhouse gas and air quality community and, hence, such methods have largely been replaced by quantitative inverse modelling. Although, some of these shortcomings may be addressed solely by text modifications and additions, others may require revisiting some of the analysis.

Thank you for your thorough reading of the manuscript. Your expert comments helped us to better frame the scope of our study, and to increase the precision of the paper's intention.
The suggested analysis also contributed to improve the study by adding clearness to the results and exactitude to the presented methodology.

Please find our answers in the following.

Major comments

Alternative attribution methods: The introduction focuses on source attribution methods that have been used for arctic aerosols and/or ice nucleation particles (L43f). However, similar methods have been applied to long-lived and short-lived compounds for more than three decades. In general, three types of statistical source attribution methods can be distinguished based on different kinds of Lagrangian modelling: 1) the present ratio method, which traditionally was termed 'potential source contribution function' (PSCF) or 'Ashbaugh method', 2) concentration-weighted trajectory (CWT) or 'trajectory statistics' and 3) inverse modelling. A fourth, even more fundamental method, would be the interpretation of back-trajectories or footprints (or potential emission sensitivity (PES), as used by the authors) of individual or aggregated events/samples. The three statistical methods are described/tested in the manuscript by Fang et al. (2018), which is used for motivation also in the current manuscript. Another example/review of these methods is by Brunner et al. (2012) and gives additional references to the origin of the methods. Although, both studies focus on long-lived greenhouse gases, these alternative methods need to be mentioned in the motivation. Of the abovementioned methods, 1 and 2 are purely qualitative methods that can only provide indications of potential source areas but cannot determine actual emission strengths. Both methods can be applied to output from single trajectory

models and from Lagrangian particle dispersion models (LPDMs, see next comment). Method 3 (inverse modelling) is a quantitative method for source attribution (location and strength) and should be the tool of choice for greenhouse gas emission attribution, but also applied for air pollutants. The manuscript wrongly states that Fang et al. (2018) 'do not present a protocol of use capable to identify sources of the studies atmospheric species' (L51f). However, Fang et al. clearly encourage the use of inverse modelling over the two other methods. Furthermore, on L501 it is stated that 'standard ratio method is more advanced than widely used inverse modeling analysis'. That's exactly not the case, as explained above. Inverse modeling is (usually) guided by prior information and through its quantitative nature and rigorous uncertainty treatment superior to the PSCF or CWT methods. This does not mean that testing and applying the PSCF method as in this manuscript should not be attempted, but the introduction and conclusion need to reflect the state of research in source attribution methods more completely.

Reply: Thank you for this comment that pinpoints a misrepresentation of the diversity of source identification methods in the introduction. Indeed, the motivations of the present paper were not clear in the original version and we missed some discussion about the existing literature. We have corrected these two issues by adding a better presentation of the literature, and by noting that the paper aims to propose and evaluate a low computational cost source identification method, with little to no a priori knowledge on emission sources. This is done in order to pinpoint the low reliability of the interpretation of raw backtrajectories and PSCFs, as they are often seen in the literature.

Based on this comment, and on a deepened literature review, we modified the introduction from L45 as follows:

"Even though the methods used vary from a study to another, the conclusions about the possible emission sources are often interpreted similarly among the community, which could be misleading.

[...]

Four different approaches that use Lagrangian modeling can be cited: 1) inverse modeling methods, 2) ratio methods or potential source contribution function (PSCF), 3) concentration-weighted trajectory (CWT) methods, and 4) direct interpretation of single or aggregated backtrajectories. Inverse modeling (1) methods have been intensively used and developed since the 2000s, particularly for the retrieval of greenhouse gas emissions (Stohl et al., 2009; Manning et al., 2011; Brunner et al., 2012; Fang et al.,2015). They are more rarely applied to aerosol emissions source identification because of the challenge in their high temporal and spatial variability (Dubovik et al., 2008, Partridge et al., 2011). Furthermore, these methods generally rely on a priori input of the emissions, such as satellite observations or existing but imprecise emission inventories. PSCF (2) (Ashbaugh et al., 1985; Zeng and Hopke, 1989) and CWT (3) (Hsu et al., 2003) are statistical methods that rely entirely on backtrajectory modeling and measurement time series. Because they are both easy to set up, they are intensively used (e.g. Polissar et al., 1999; Hirdman et al., 2010; Irish et al., 2019; Ren et al., 2021). Nevertheless, interpretation of raw backtrajectories (4) is still common in the literature (references of INP studies herein above), and can lead to spurious conclusions about emission sources. Fang et al. (2018) evaluated the performances of inverse modeling against CWT's and PSCF's, and concluded that the high quantitative power of inverse modeling surpasses the qualitative results of CWT and PSCF. Yet, PSCF and CWT are computationally low cost and can give useful insights when correctly applied and interpreted. The direct interpretation of backtrajectories remains the less reliable approach."

Finally, to make the ambitions of the study more clear, we modified the first sentence of the paragraph L54 as follows:
"The present study focuses on the evaluation of low computational cost methods with little to no a priori knowledge on emission sources. The atmospheric species aimed here are short-lived atmospheric compounds, such as aerosol particles, for which global observations areparticularly challenging."
On L56 we added the term *"qualitatively"* to highlight the expectation for the method we evaluate and use:
"[...] in order to assess its ability to ***qualitatively*** *retrieve known sources [...]"*

Single trajectory vs. Lagrangian particle dispersion models: The authors seem to treat studies carried out with (single) trajectory models and LPDMs synonymously. However, there are important differences to be considered that are not well reflected in the motivation and placement given in the present manuscript. Single trajectory methods do not reflect the dispersion of air masses at all and, hence, do not well reflect quantitative air mass transport, especially when transport in the atmospheric boundary layer is concerned. Lagrangian dispersion models describe turbulent (and convective) transport through a stochastic process on a multitude of model particles (air parcels). Applied correctly, they provide quantitative transport statistics. Consequently, source attribution methods solely based on single trajectory models suffer even more non-quantitative challenges. To make the distinction between the two types of Lagrangian models I suggest introducing them as such in the introduction and to refer to output of these models by back-trajectories and PES, respectively.

Reply: Thank you for picking this lack of accuracy in the wording used in the manuscript, and for the suggestion. We mixed references to literature that use either single backtrajectories or dispersion models in order to illustrate the variety of existing methods. As major comment 1 highlighted — as well as specific comment 3 —, both the scope of the cited studies and the method approached were mixed, as a result the overview of the existing methodologies was unclear.
We kept the mixed list of references on INP sources identification to illustrate variety of approaches used for these specific studies, and have changed the discussion to include a distinction regarding the type of analysis used (single backtrajectory interpretation and PSCF):
"For example, this kind of analysis is increasingly used and presented alongside the analysis of INP observations. Some studies use straightforward interpretation of single or dispersed backtrajectories (Allen et al. ,2021, Hartmann et al., 2020, Hartmann et al., 2021, Porter et al., 2022, Wex et al., 2019, Yun et al., 2022), while others performed more advanced analysis like potential source contribution functions (PSCFs) (Irish et al., 2019, Si et al., 2019)."

In order to remove the ambiguity about single backtrajectories and particle dispersion modeling, we moved the presentation of Lagrangian models from Sect. 2 to the introduction (as suggested in specific comment 5), and we added the following definition:
"In the following, we will refer to single backtrajectory model output as backtrajectories, and to dispersion model outputs as potential emission sensitivity (PES). Both model types will be referred to as Lagrangian models."

Improved ratio method: It would be interesting to see how the different modifications of the old ratio method impact the results of the improved method. Is the main improvement through the cropping of

low residence time areas, the background subtraction or the additional modification through the low concentration footprint? From Fig. 3 it appears as if not too many of the selected observations for the high concentration case change from the original selection to the above-background selection. If the residence time cropping is the main cause of improvement, the conclusions drawn here may be very specific to the present setup of observing sites and source areas.

Reply: Thank you for this suggestion of analysis. We ran the analysis and we added a dedicated paragraph in Sect. 3.4 *Results comparison*.

In order to compare the contributions of the three modifications to the standard ratio method (filter of low FPES, background subtraction and composite ratio), we used the averaged success level of detection for the three tracers and the five stations (average of the fifteen values of detection level). The average value of the standard ratio method is 0.9. The improved ratio method (with the three modifications), reaches an average of 2.0.
To test individually each modification, we ran the standard ratio in combination with one or two modifications of the improved ratio. We present the result in a new table, reproduced below. The values correspond to the averaged success level when the column and row modifications are used. Thus, the diagonal values refer to the test when a single modification is used, the others when both the row and column modifications are used.

	FPES filtering	background subtraction	composite ratio
FPES filtering	1.3	1.5	1.6
background subtraction	1.5	1.0	1.1
composite ratio	1.6	1.1	1.1

The values are rounded at the first two digits.

The highest improvement is due to the FPES filtering, but without the background subtraction or the composite ratio, the averaged detection score never exceeds 1.6. The three modifications are needed to reach an averaged detection level of 2, which was defined as a successful detection (Sect. 3.1).

In the text, the new paragraph is written as follows:

“ In order to assess the contributions of the three modifications (FPES filter, background subtraction, composite ratio) to the standard ratio introduced with the improved ratio, the source identification has been run in the standard ratio setup with one or two improved ratio's modifications. The results are presented in Tab. 4. Each cell value corresponds to the averaged success level when the column and row modifications are used. Thus, the diagonal values refer to the detection score when a single modification is used in addition to the standard ratio; the others, when both the row and column modifications are used.

The highest improvement is due to the FPES filtering, with an improvement of 0.4 compared to the standard ratio. The composite ratio allows for a 0.2 rise, where the background subtraction leads to an improvement of only 0.1. The best combination appears to be the composite ratio with the FPES filter, which improves the score of the standard method by 0.7. The combination of the composite ratio and the background subtraction only account for a 0.2 improvement. Even though the

latter seems to be a small improvement, the full potential of the improved ratio method is only reached when the three modifications are used all together.

In other words, the three modifications are needed to reach an averaged detection level of 2, which is the threshold of a confident successful detection (Sect. 3.1).”

We moved Tab. 4 of the submitted manuscript to the appendix because its information were redundant with Fig. 4, but we kept it because it helps to compare the standard ratio and the improved ratio scores.
We also added, in the supplementary materials, the detailed performances (for the three tracers and fives stations) of the six combination of modifications tested here, with the same bar plot presentation than Fig. 4.

Specific comments

L8, 'commonly used back-trajectory analysis': Not sufficiently specific. If by back-trajectory, single trajectory simulations without dispersion are targeted, I would agree. Otherwise, this is too general and needs to be more specific to the kind of analysis tested in the manuscript.

Reply: By ‘commonly’ we meant three analysis methods seen in the literature: direct interpretation of the backtrajectories (whether they are single or dispersed backtrajectories), concentration-weighted backtrajectories (CWT), and PSCFs. However, we agree that “commonly” is not a specific enough term. We made the following change on L8:
“*The results show that direct interpretation of backtrajectories, and potential source contribution functions (PSCFs) are often unreliable in identifying emissions sources.”*

L35f: There are references given for the first kind of studies mention in the sentence, but not for the second kind (correlation with chemical tracers). Please provide examples.

Reply: We added references for the use of correlation with chemical tracers: **Jiang et al. (2009)** about carbon monoxide from biomass burning and fossil fuel combustion, and **Park et al. (2017)** about dimethyl sulfide from phytoplankton blooms.

L42f: References given for INP studies use a mix of different methods, from single trajectories to ratio methods relying on LPDM output. These different kinds should be listed separately and references given accordingly. The publication by An et al. (2014) does not seem to contain INP at all, but focuses on CO. Why was it mentioned?

Reply: We have modified the sentence by introducing a categorization of the papers according to the method they use, as suggested. The paragraph is now as follows:
“For example, this kind of analysis is increasingly used and presented alongside the analysis of INP observations. Some studies use straightforward interpretation of single backtrajectories or LPDM

outputs (Allen et al. ,2021, Hartmann et al., 2020, Hartmann et al., 2021, Porter et al., 2022, Wex et al., 2019, Yun et al., 2022), while others performed more advanced analysis like potential source contribution functions (PSCFs) (Irish et al., 2019, Si et al., 2019).”

Indeed, An et al. (2014) was mentioned among the INP studies by mistake. The confusion came from the fact we wanted it to be cited for its use of FLEXPART. Thank you for pointing this. We have removed the citation.

L49ff: The interpretation of Fang et al. (2018) was already mentioned in the general comment. The final sentence of the paragraph needs to be more specific again as OSSEs are routinely carried out for inverse modelling of greenhouse gases.

Reply: We modified this part of the introduction, which corrects the interpretation of Fang et al. (2018) presented in the paper. See answer to major comment 1.

L65ff: The paragraph introducing the specific analysis method should be moved into the introduction, as it contains general discussion of modelling concepts and is used to motivate the present study and not to describe methodological details.

Reply: This was addressed in response to major comment 2, we followed this suggestion by moving the corresponding paragraph up to the introduction. It helped to clarify the distinction between the single trajectory models and the dispersion models.

L68f: The list of Lagrangian models should be sorted by model type, single trajectory models (HYSPLIT, LAGRANTO) vs. LPDMs (FLEXPART, STILT). HYSPLIT can be run in dispersion mode as well. Another frequently used LPDM is NAME (Jones et al., 2007).

Reply: Thank you for the suggestion, this classification helps to make the text more clear. We also appreciate the reference of the NAME model. We modified the text accordingly; it is implemented in the introduction as suggested in the previous comment.

L71: Here is an example of the use of back-trajectories for the output of a LPDM. As mentioned in the general comment above, replace it with potential emission sensitivities.

Reply: We transformed the introduction of the corresponding section where this sentence belonged by following major comment 2, therefore it has been completely changed. However, we have been careful to use the correct nomenclature when talking about LPDM outputs. Specifically, we replaced ‘backtrajectory’ by ‘backward’ where appropriate, and used PES when needed.

L93: What is meant by 'assimilated' in this sentence? Did you mean associated? I think corresponds would work best for all tracer types listed in the sentence.

Reply: ‘Assimilated’ was used here with the meaning of ‘being analogous’. To remove ambiguity avoiding repetition, we modified the sentence using ‘correspond’, ‘represent’, and ‘ associated’:

"The continental tracer corresponds to mineral dust or continental biogenic aerosols, the open ocean tracer represents sea spray emissions, and the sea ice tracer is associated with blowing snow emissions."

L98f: If the simulated tracers are supposed to represent aerosols, why was an exponential decay chosen over a tracer undergoing typical aerosol removal processes (settling, dry and wet deposition)? The exponential decay somewhat simplifies the behaviour of a real tracer and may represent an easier target for the source attribution than real aerosol. Please give additional motivation for this choice and discuss the limitations.

Reply: In the following text, starting at L212, we refer to the modeled tracers as 'idealized tracers'. From L229 to the end of the manuscript, we call the assessment protocol an 'idealized situation' or 'idealized tracer experiment'. This refers to the fact that the tracers are on purpose made as simple as possible in order to keep away uncertainties indirectly related to the detection method, namely parameterization of dry and wet removal in FLEXPART. The tracers needed a removal mechanism because of the long duration of the modeling experiment, but in order to keep the study as general as possible, we decided not to set advanced removal processes since those highly depend on the nature of the studied species, even among aerosols because of their size and hygroscopy (Ohata et al. 2016, Farmer et al. 2021). The exponential decay was a way of having the exact same removal in both the tracer mechanism used in WRF-Chem and FLEXPART.

The main limitation is that our evaluation can not account for the uncertainties on removal processes. As discussed above, it enables keeping the evaluation as general as possible, since it focuses on short-lived atmospheric species.

In order to detail this important point of the methodology, we added a paragraph that properly presents the experiment as idealized (which was missing), and discusses the motivations for that setup:
"Because of the long duration of the modeling experiment, the tracers would accumulate infinitely in the domain without removal. In order to keep the study as general as possible, we decided not to set advanced removal processes, namely dry and wet removal, since those highly depend on the nature of the studied species. This is the case for aerosol particles, whose removal strongly depends on their size and hygroscopy (Ohata et al. 2016, Farmer et al. 2021).
Since the study focuses on short-lived species, the tracers are removed thanks to an exponential decay with a characteristic time of three days. This allows the exact same removal in both WRF-Chem forward and FLEXPART-WRF backward simulations, thus the evaluation is free of the uncertainties on removal parameterizations.
In that way, the evaluation setup is idealized, and accounts only for the best performances that can be expected from the tested methods."

We added a discussion on the related limitation in Sect. 6 (Conclusions), after the discussion on the back-tracking duration (L482):
"A related point is the setup of removal processes for the evaluation experiment. A removal by exponential decay was used to represent short-lived species. This causes two important limitations: 1) the uncertainties on removal processes are not taken into account in the results, 2) the present evaluation does not explore the effects of different removal processes on the performance of the

method. Consequently, one should pay special attention to what removal parameterization is set in the LPDM when attempting an emission source identification."

L110f: Does this mean that wet deposition was considered in the simulated tracers after all? Please clarify.

Reply: This was an imprecision. Wet removal was not implemented in the removal of the tracers in FLEXPART or in WRF-Chem either. We removed this sentence.

L140f: 'particle diffusion all along the edges'. Sounds odd. How an LPDM simulates atmospheric dispersion was mentioned above. If LPDMs are introduced properly in the introduction, FLEXPART can just be introduced as such and its output being PES.

Reply: This sentence brought confusion. After adding some description about LPDMs and the nomenclature on backtrajectories and PES in the introduction, this sentence became useless. We have removed it.

L145f: Does this mean with a mass proportional to the WRF-simulated concentrations? In backward mode the mass given in a FLEXPART RELEASE is ignored and set to unity. Hence, such an approach would not, and should not, reflect the observed concentrations at the receptor at all.

Reply: The tracers are not emitted with a mass proportional to the WRF-Chem concentration, and are indeed ignored and set to unity.

L147: The approach to release/initialise model particles/trajectories in a larger area is usually not applied to LPDMs that can simulate dispersion. This is usually done when employing single-trajectory models to mimic dispersion by an ensemble of trajectories with different initial locations. It is one big benefit of LPDMs that they can be used to treat point sources properly and in backward mode this means that they can correctly represent point measurements, better than Eulerian models, where you need to interpolate to a given location.

Reply: Thank you for this remark. Indeed, this is unfortunate we did not take advantage of this LPDM property of FLEXPART.

However, we ran a test for a single day in order to assess the sensibility to the release box size. The difference in the FPES values appears to be low. Although in some grid cells very localized around the emission coordinates, the FPES difference can reach several hundreds of seconds. The average of the difference over the whole domain is around 10^{-2} s, while the corresponding average FPES value is around 1 s. Visually, the difference affects only the edges of the FPES plume, where the values are 3 orders of magnitude lower than the averaged signal.

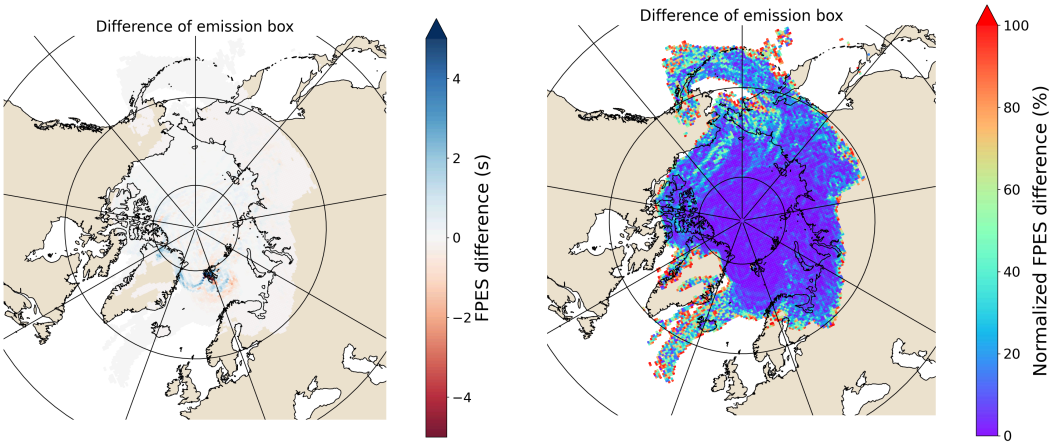

Figure. Difference between the FPES of two FLEXPART runs: one emits particles in a box of 50km×50km×10m, the other in a box of 10m×10m×10m. The left panel shows the absolute difference in seconds; the right panel shows the difference normalized by the FPES values.

L148: 10'000 particles per day sounds a bit low to produce robust FLEXPART simulations. Usually, release rates of ten thousand of particles per hour are used. If I interpret Irish et al. (2019) correctly, they even released 100'000 every 20 minutes. Raut et al. (2017) seem to have released 10'000 every ten minutes along their flight track.

Reply: The manuscript indicated the wrong number of emitted particles. 100’000 have actually been used.

L149f: This should not happen when a sufficiently larger number of model particles was selected. The residence time calculated by FLEXPART should not be proportional to the number of released particles, because each air parcel in FLEXPART is assigned an equal time fraction that is proportional to 1/N, N being the number of released particles. I suggest performing two test simulations for a single day and set the number of released particles to something like 240'000 and 480'000 particles per day and compare the output of such runs with each other and the previous run with 10'000 particles. If the two with the large number show less difference between each other than compared to the one with 10'000 particles, I would think there is a strong need for a note of caution to be given for the employed FLEXPART setup.

Reply: That is true, this sentence was mistaken about the behaviour of the PES regarding the number of emitted particles. We removed it.

We did the suggested test to compare the FPES outputs of four FLEXPART runs: with 100 k, 240 k, 480 k and 960 k particles emitted. The observed difference are the most important between 100 k and 240 k emitted particles (comparatively to the differences 480k - 240 k and 960 k - 480 k). The difference decreases when the number of emitted particles increases (i.e. difference for 960 k - 480 k is smaller than difference for 480k - 240 k, which is smaller than 240 k - 100 k). However, the observed differences are one to three orders of magnitude lower than the FPES in the corresponding grid cells, and affect only some specific areas, which are small compared to the domain size. Quantitatively, 0.2 % of the cells covered by the FPES plumes are concerned by a difference.

In order to account for these tests, and to give a recommendation on the emission rate, we modified the sentence L149 as follows:
"The sensitivity of the PES plume to the number of released particles is inversely proportional to this number. Although this sensitivity is low, a rate of 10,000 particles emitted per hour is recommended."

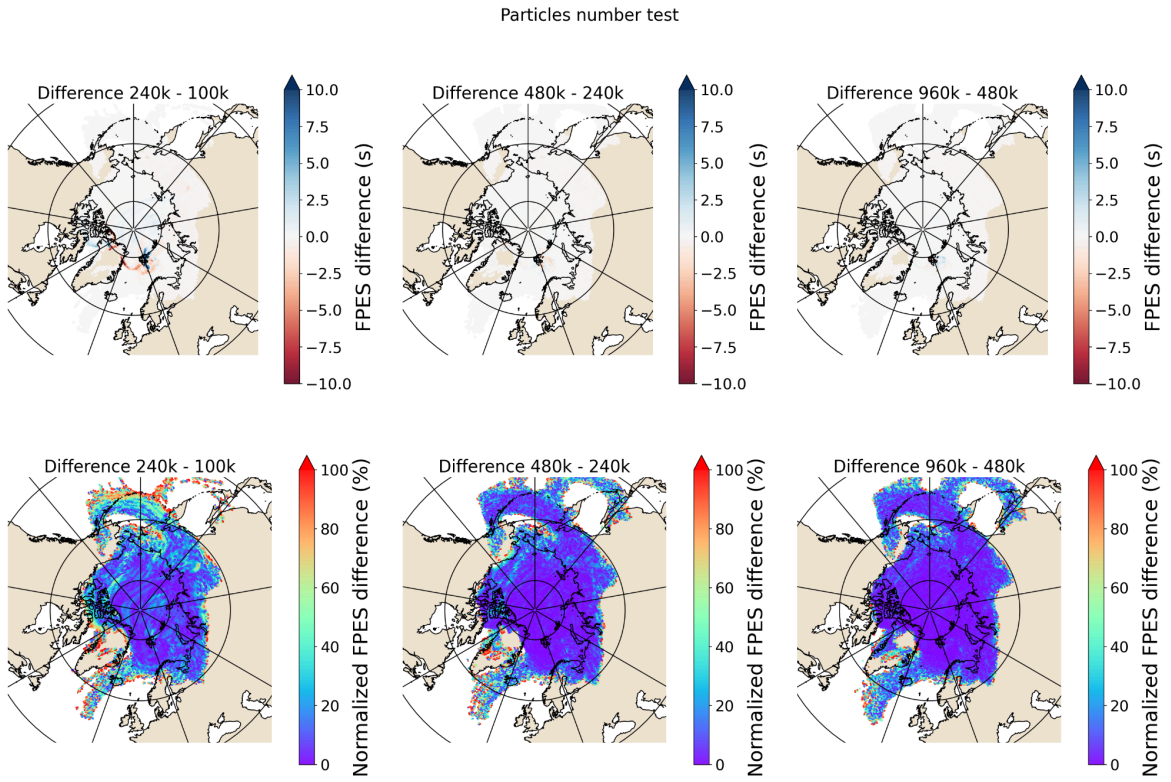

Figure. Difference between different FLEXPART runs set up with various numbers of emitted particles. The upper panel shows the absolute differences expressed in seconds; the lower panel shows the difference normalized by the FPES values.

L186: Consider '… no longer the presence but the absence of sources or even the presence of sinks'.

Reply: Thank you for the suggestion, we adopted it in the text.

L197ff: Is this really a fair evaluation? Since source strength is not uniform over the area but depends on wind speeds, a better comparison would be the average flux map used in the WRF simulations.

Reply: We agree that identifying the land surface type is a less strict requirement than identifying the precise origin of the flux on the correct land type. But the analysis shows that this level of precision is already challenging to achieve. This choice for the evaluation protocol is in accordance with the idea of 'idealized' experiment mentioned above and in the text.

L201: Not quite clear how D is calculated. An equation or additional description would be helpful.

Reply: D is the detected contribution fraction ($D \in [0, 1]$) of the land type from which the tracer has been emitted. The corresponding equation is $D_T = R_T / R$, with T referring to the considered tracer

(sea ice, open ocean, continent), and R is either the standard or improved ratio. R_T is the sum of the values of R over the surface type corresponding to the tracer T.

For clarity, we followed your suggestion and added the equation to the text.

In addition, we spotted that the success level of detection was miscalled D in the caption of Fig. 4, which added to the confusion. We corrected the text consequently.

L225: For the reader to understand the magnitude of the failure, it would be good to show such a failed experiment as well, similar to Fig 2.

Reply: We added, in the supplementary materials, all the maps for the standard (Fig. S1) and for the improved method (Fig. S2). Figure S1 should help the reader to understand what a fail detection looks like, and Fig. S2 illustrates the discussion about the results of the improved method (c.f. comment about L291). These figures are referenced in the main text. Both figures' results are already summarized in Fig. 4.

L229ff: I don't think it is the uncertainty of the transport model that is the main reason for failure here. I would rather think that it is the relative position of source areas and receptors. Any source attribution method based on a single observing site will suffer from a so-called shadowing effect. This tends to falsely assign emissions to areas that are upwind of the true emitting area. In the present case and for the sea ice and ocean tracer, the continental areas are mostly in such a configuration. The only way to robustly overcome this problem, is a network of sites that can observe gradients across the domain or at least 'observe' the same source area under different flow directions (as demonstrated in section 3.5).

Reply: Thank you for this very clear and concise description of the shadowing effect. We hope that it is ok if we include it as-is in the text:

" We can identify three reasons for this behaviour [...]

Secondly, and to a lesser extent [...]

Last but not least, any source attribution method based on a single observing site will suffer from a so-called "shadowing" effect. This tends to falsely assign emissions to areas that are upwind of the true emitting area. In the present case and for the sea ice and ocean tracer, the continental areas are mostly in such a configuration. The only way to robustly overcome this problem, is a network of sites that can observe gradients across the domain or at least 'observe' the same source area under different flow directions (as demonstrated in Sect. 3.5)."

L250: Usually this is called background subtraction, which is supposed to separate slowly varying background concentrations from recently added pollution events (e.g., Ruckstuhl et al. 2012, Resovsky et al. 2021). How was this done here?

Reply: We edited the text to explain our method, and added the suggested references.
"This is similar to the background subtraction method of Ruckstuhl et al. (2012) and Resovsky et al. (2021). Practically, we estimated the background seasonal concentration by smoothing the series with a locally weighted scatter plot smoothing (LOWESS) filter, and then subtracting the background trend from the concentrations to keep the high frequency signal from recently added emission events."

L251: Using a residence time threshold to remove 'boundary' effects has been done in many source attribution methods before. For example see the factor W_ij in Fang et al. (2018) that is based on the number of trajectories passing through a grid cell, but is nothing else than a residence time threshold when translated to LPDM output.

Reply: We included this discussion and the reference: *"Concerning the over representation of the continental area, a cutting threshold on FPES is set in order to filter the less significant backtrajectories. This was not done in the implementation of Hirdman et al. (2010), but is similar to the approach of Fang et al. (2018), who suggested excluding grid cells crossed by too few trajectories."*

L265: 'leaks' sounds a bit strange in this context. I would prefer the term 'shadowing' or more general 'caveats'.

Reply: Thank you for informing us of the proper terminology. We updated the text:
*"presenting an overflowing **"shadowing"** of the plume on the continental regions **(see section 3.2)**"*
and L292:
*"This shows that **shadowing** can still happen"*

Equation 6: This is the innovative part of the method and it should be mentioned as such.

Reply: We stressed this in the text: *"Additionally, unlike the ratio of high concentrations, we aim to select as many regions as possible that are detected as unlikely sources. **To our knowledge, taking into account these areas of low concentrations in a PSCF method has not been tried in the past and is the most innovative part of our method.**"*

And in the conclusions: *"With the improved ratio method, **we created a novel approach by introducing a composite ratio** that takes advantage of the information contained in the detection signal associated with both the highest and the lowest measurements."*

L274: What is Rl_uparrow_33 set to elsewhere? In areas where it does not take the value of Rl_33.

Reply: Rl_uparrow_33 is the value of Rl_33 when Rl_33 is above 0.33. The threshold of 0.33 refers to the threshold that defines the low concentrations, and corresponds to the ratio values that are significant (as detailed in Sect. 2.3, L184-186).

L278: What 'bars'? Isn't this simply the number of correct detections/attributions?

Reply: L278 we replaced "bars" by "number of correct attributions".

L291ff: It seems that this paragraph describes figures that are not shown in the manuscript. Consider adding to a supplement or annex.

Reply: We included the maps referenced in the text as a supplementary figure. We added the reference to it in the text. For the sake of comparison, we also added in the supplementary materials, the corresponding figure for the standard ratio method.

L300f: A more suitable example in this context would be the networks for greenhouse gas observations that are used to attribute global and regional emissions/sinks (e.g., ICOS, WMO GAW, NOAA flask network, AGAGE).

Reply: We added the examples: *"Some observations of atmospheric species are made within a measurement network composed of multiple experimental sites (e.g. ICOS, WMO GAW, NOAA flask network, AGAGE)."*

L320ff: Again, a reference to existing observing networks that are used for source attribution would make a lot of sense.

Reply: We modified the text *"Therefore, the present work encourages the development of observational networks or coordinated field experiments dedicated to identify short-lived species at the high latitudes, following the example of the networks for greenhouse gas observations that are used to attribute global and regional emissions."*

L324: I suppose with 'basic back-trajectory analysis' you are referring to what I called the fourth (non-statistical) method, that only looks at trajectories/footprints of individual observations. I don't quite understand how such an approach is reflected in Fig. 6.

Reply: We indeed meant looking at the footprints of observations without statistical analysis. For clarification, we replaced 'basic back-trajectory analysis' with "analysis of raw FPES". This is shown on Figure 6b. We agree that this is not looking at an individual observation, but rather an aggregate of all the high-observation FPES that could be looked at individually. We still think it is a good illustration of the quality of the results that can be obtained with this 4th method, and that is somehow similar to the aggregation of single backtrajectories as seen in several studies (e.g. Hartmann et al., 2021, Porter et al., 2022).

Furthermore, we also modified the sentenceL327:
*"These representations are often used for **quick qualitative source identification** in studies using backtrajectories **or PES** [...]"*

L362f: Is the selection of one observation every two days enforced onto the already cropped, one-year data set or on the two-year data set? If the latter, then it is surprising that a decrease in performance is observed, whereas none was seen when shortening the time series from two to one year. Both cases should have the same number of total observations in the analysis.

Reply: The selection of one observation every two days has been set over the full two-year dataset. So, indeed the number of points is the same for the daily dataset over one year and the semi-daily two-year dataset (i.e. 365 concentration values).
An explanation for the observed difference in the results might be that reducing the frequency causes missing some variability of the emission sources. The performance of the detection would therefore

not only depend on the total number of points, but also on what variability is captured, as suggested by the results.

L364: This is a trivial conclusion. More observations should always improve this kind of statistical source attribution analysis. However, there is also a limit to enhancing temporal resolution, since atmospheric variables are usually auto-correlated and the amount of independent information cannot be increased by measuring more frequently. However, this is for time scales shorter than a day and is probably not what was referred to here.

Reply: We agree, this is not a very surprising conclusion for an experienced reader, but we still think that it should be reminded for those less familiar with these methods. Especially because it can be tempting to apply the method on datasets with low time resolutions (i.e. lower than every other day).

L370: Consider 'arbitrary results' instead of 'insignificant information'.

Reply: Thank you for the suggestion; we adopted it.

L373f: Exactly, that's why filtering for a remote source may be dangerous and the numbers obtained here cannot easily be generalised for other sites or compounds.

Reply: We agree with the reviewer, and to clarify this we added a short discussion on this effect in the conclusion when discussing filtering:
"*We introduced a filter on the lowest backtrajectory values in order to get rid of the less statistically significant ones. It led to a better representation of the three surface types, which resulted in a dramatic improvement of the detection results. The effect of this filtering is to shrink the result of the method close to the measurement station, and it is possible that the parameters we used might not be generalized for other regions or compounds, although the methodology to get the best filtering level can be generalized.*"

L407: Gilardoni et al. (2019). There is no source attribution analysis presented in this document.

Reply: We cited Gilardoni et al. (2019) because of this sentence: "BC atmospheric concentration in the Arctic region is controlled by BC emissions at high and middle latitudes".
However, we agree, since it is not a peer reviewed article but a report, it is a poor reference, and we replaced it by a primary source reference: Xu et al. (2017), *Source attribution of Arctic black carbon constrained by aircraft and surface measurements*.

L422: Panel a of Fig. 9 does not show a ratio map.

Reply: Indeed, the sentence refers to the panel (b). We did the correction. Thank you.

L427: How does the potentially long transport time from the Caspian Sea agree with atmospheric lifetimes of MSA? Would we not expect MSA to be mostly destroyed?

Reply: The lifetime of MSA in the atmosphere is largely driven by the atmospheric conditions and can reach several weeks (Mungall et al., 2018). As a result, long-range transport is possible and has been discussed for a range of DMS oxidation products at another remote station, the Bolivian Andes (Scholz et al., 2023). Furthermore, long-range transport as source of Arctic aerosols originating from

Central Asia is a phenomenon observed in the past (Marelle et al., 2015). As no studies so far reported DMS or MSA emission from the Caspian Sea region, it might be speculative to conclude on a contribution of this region to the MSA observed at Zeppelin Observatory, as already mentioned in the manuscript. Nevertheless, such a contribution can also not fully be ruled out.

In order to address this point in the manuscript, we added the following:
"Such long range transport is surprising but not impossible: long-range transport of aerosols to the Arctic from Central Eurasia has been observed in the past (Marelle et al., 2015) and the typical lifetime of aerosol MSA against OH oxidation is a few weeks (Mungall et al., 2018)."

L468: Repeated from comment above. It's probably more the shadowing effect than the continental dominance.

Reply: In order to account for the shadowing effect in this sentence, we rephrased it as follows:
"The results of the standard ratio method are influenced by the geographical and layout wind configuration, causing an over-representation of the continental areas and a shadowing effect in the detected sources."

L501: 'standard ratio method is more advanced than widely used inverse modeling analysis'. This is certainly not true, unless you wanted to express that ratio methods were the most frequently applied tool for analysing INP sources in the Arctic.

Reply: You are right, here we misused "*inverse modeling*". We wanted to refer to studies that use backtrajectory modelling approaches for emission sources identification, with little or no post-processing analysis.
In order to dispel doubt, we clarified this sentence as follows: "*Although this standard ratio method is more advanced than most of the backtrajectory analysis used in studies about short-lived atmospheric species [...]"*

We also replaced "*inverse*" and "*reverse modelling*" by "*backward modelling*" everywhere the confusion was possible.

Technical comments

L27: '.' missing after (Matus and L'Ecuyer, 2017).

Reply: We missed that one. Thank you.

The bibliography does not comply with the Copernicus style.

Reply: Thank you for pointing this out. We used the Copernicus LaTeX style file to create the bibliography, but if something went wrong, we will fix it with the editor in the production stage.

References

Brunner, D., Henne, S., Keller, C. A., Vollmer, M. K., and Reimann, S.: Estimating European Halocarbon Emissions Using Lagrangian Backward Transport Modeling and in Situ Measurements at the Jungfraujoch High-Alpine Site, in: Lagrangian Modeling of the Atmosphere, edited by: Lin, J. C., Gerbig, C., Brunner, D., Stohl, A., Luhar, A., and Webley, P., Geophysical Monographs, AGU, Washington, DC, 207-221, doi: 10.1029/2012gm001258, 2013.

Fang, X., Saito, T., Park, S., Li, S., Yokouchi, Y., and Prinn, R. G.: Performance of Back-Trajectory Statistical Methods and Inverse Modeling Method in Locating Emission Sources, ACS Earth and Space Chemistry, 2, 843-851, doi: 10.1029/2012gm001258, 2018.

Jones, A., Thomson, D., Hort, M., and Devenish, B.: The U.K. Met Office's Next-Generation Atmospheric Dispersion Model, NAME III, Boston, MA, 2007, 10.1007/978-0-387-68854-1_62, 580-589, doi: 10.1007/978-0-387-68854-1_62, 2007.

Resovsky, A., Ramonet, M., Rivier, L., Tarniewicz, J., Ciais, P., Steinbacher, M., Mammarella, I., Mölder, M., Heliasz, M., Kubistin, D., Lindauer, M., Müller-Williams, J., Conil, S., and Engelen, R.: An algorithm to detect non-background signals in greenhouse gas time series from European tall tower and mountain stations, Atmos. Meas. Tech., 14, 6119-6135, doi: 10.5194/amt-14-6119-2021, 2021.

Ruckstuhl, A. F., Henne, S., Reimann, S., Steinbacher, M., Vollmer, M. K., O'Doherty, S., Buchmann, B., and Hueglin, C.: Robust extraction of baseline signal of atmospheric trace species using local regression, Atmos. Meas. Tech., 5, 2613-2624, doi: 10.5194/amt-5-2613-2012, 2012.

The references used in the answers are given in the revised manuscript.