Response to the reviews of manuscript amt-2021-429: “Automated identification of local contamination in remote atmospheric composition time series”, by Ivo Beck, Hélène Angot, Andrea Baccarini, Lubna Dada, Lauriane Quéléver, Tuija Jokinen, Tiia Laurila, Markus Lampimäki, Nicolas Bukowiecki, Matthew Boyer, Xianda Gong, Martin Gysel-Beer, Tuukka Petäjä, Jian Wang, and Julia Schmale to *Atmospheric Measurement Techniques*

Questions from the reviewers are written in *blue*, our answers in black, text copied from the manuscript is written in *italic*, and all changes in the manuscript are typed in *italic red*. When referencing page and line numbers, we are always referring to the old version of the manuscript.

Answers to Reviewer 1 are from p. 2 to 13, to Reviewer 2 from p. 14 to38, references start on p. 38.

Reviewer 1:

This work describes an algorithm able to distinguish between the local component and the anthropogenic contamination of environmental related datasets. The algorithm is shared not only as a flow diagram but also as a functional code, which makes it even more important. Data cleaning is a cumbersome procedure that anyone working with environmental monitoring has faced. Therefore, this work can contribute towards the automation of these labour intensive procedures.

Most articles related to data cleaning deal with a specific dataset only, and the reader is always left wondering what are the limitations of the proposed algorithm and whether it would be worthwhile to apply it to other datasets. In this work, the same algorithm is applied to several different instruments to prove its wide applicability. I must note though that practically two different methods are presented in Section 2.4.1 that have been incorporated under one software.

We thank the reviewer for this positive and constructive review. Below we answer all comments in detail and believe that the manuscript has greatly improved thanks to the comments.

R1.1

The only weakness of the proposed method is that the user should decide on up to 7 parameters to make the code operate optimally, even though it is discussed in the manuscript that 3 are necessary and the remaining optional. This makes the proposed algorithm quite subjective. Can the authors comment on that?

The reviewer is correct that setting between 3 and 7 parameters is not entirely objective. When we refer to "objective", we meant that the treatment of the dataset in itself is consistent, because these parameters are not changed. All parameters are set once in the beginning and applied to the whole dataset. Hence the algorithm is more time efficient and more consistent in itself in terms of strictness, compared to, for instance, a manual cleaning method, where the expert has to decide on each individual data point. The 3 to 7 parameters also make the algorithm flexible. We aim to provide an algorithm, which can be used with many different remote atmospheric datasets. This comes with the trade-off that the user needs to adjust all parameters. In other words: If we restrict the parameters, we exclude more datasets and more use cases, hence the algorithm could not be applied to many other user's datasets. To emphasize the advantages of the PDA and to clarify what is meant by "objective" we have made the following change in the manuscript in section 1, at line 109:

This makes the algorithm an efficient and consistent way to detect local contamination in large remote atmospheric time series, as they exist for example from ship campaigns or remote stations. This method is objective as the treatment of the data is consistent throughout the whole time series considered, because the same value of each parameter is applied to the entire dataset.

R1.2

The manuscript is very descriptive and answers most questions that may arise. However, there are a few clarifications required, mostly related to the applicability of the algorithm (named PDA in the manuscript).

In specific

The manuscript would benefit if the underlying assumptions of applying the PDA are clearer. PDA is used only in datasets obtained in pristine conditions, where the concentration difference between anthropogenic and local components can be an order of magnitude. What is the smallest difference that can be detected? Of course this relates to the parameters selected by the user. This point should be discussed further.

This is a good point. To figure this out we would have to test the algorithm with a number of different datasets. As mentioned by the reviewer, each dataset needs its own parameters. If we published specific numbers of minimum derivatives, they would be valid only for specific datasets. However, we revised section *3.4.* and specified the underlying assumptions in more detail (see below). In Fig. A11 we also show an example where the PDA reaches its limits and a separation of contaminated from unaffected data points is difficult to make. In this example, some of the flagged data points don't exceed the "baseline" concentration at all. The difference between an unaffected and a flagged data point can be 2 cm^{-3} at concentrations of 190 cm^{-3}, or 10 cm^{-3} at 390 cm^{-3}. It is important to remember that the derivative filter threshold depends on the concentration. Distinguishing contaminated from unaffected data points in this situation is difficult in this case study, even by expert judgement. This is one of the strengths of the PDA, by tuning the algorithm to more strictness, the user can decide to be stricter. We demonstrate this in the newly added Figure A12, where we made the derivative filter stricter, with a = 0.45 (instead of 0.5) and m = 0.5 (instead of 0.55), and flagged more data points as contaminated. This leads to many more flagged data points, especially after 15:00 UTC.

We added the following sentences to section 3.4:

Another challenge for the PDA is situations where the signal is influenced by subtle contamination, which does not result in large spikes but rather in a very noisy signal with low amplitude above a background concentration. Two examples are shown in Fig. A9 and A11. These situations are also difficult to assess for an expert using the visual inspection method. The boundary between polluted and unaffected data is blurred, and the derivative filter in Fig. A11 only flags a subset of data points that protrude from the main signal. In this example, some of the flagged data points do not exceed the "baseline" concentration at all. The difference between an unaffected and a flagged data point can be 2 cm^{-3} at concentrations of 190 cm^{-3}, or 10 cm^{-3} at 390 cm^{-3} (the derivative filter threshold depends on the concentration). If we choose a stricter derivative filter, for example, with a = 0.45 (instead of 0.5) and m = 0.5 (instead of 0.55), more data points are flagged as contaminated and hence less false negatives remain (Fig. A12). However, this might also remove more unaffected data points, and it is up to the user to make this decision.

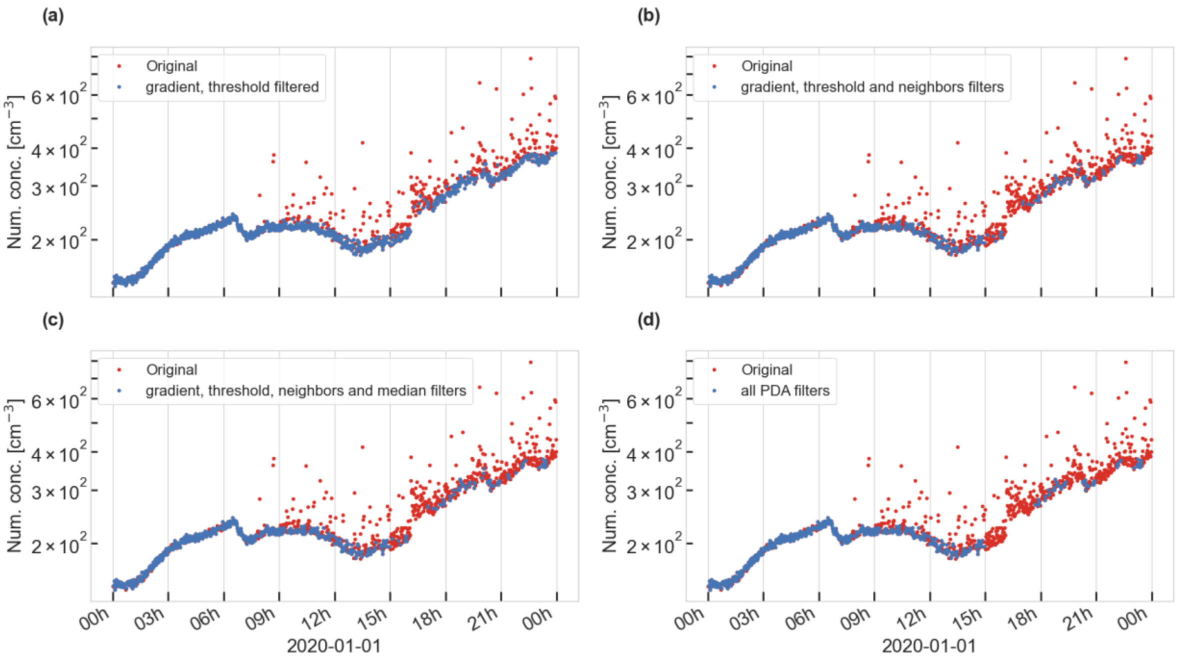

Figure A12: Same as figure A11, but with slightly stricter coefficients of the derivative filter. We chose a = 0.45 and m = 0.5 to flag more data points in this case study.

R1.3

How does the PDA respond to data gaps? Is there any restriction if gradient filter method A is applied?

How does PDA respond to the edges of the dataset? Please discuss.

How big dt in Eq 1 should be and how that is determined? How is dt related to the time resolution of the dataset and the expected duration of the anthropogenic events.

We answer above three questions together, since they all target the derivative filter. These are very important questions and it is crucial that the reader understands the underlying mechanism. We reformulated section 2.4.1 to clarify these points (see below).

We calculate the derivative using the central difference formula: $f'(x) = \frac{f(x+1)-f(x-1)}{2}$, for every data point x. This means that in case of data gaps, the derivative will be calculated over the gap, regardless of the duration of the gap. In case of large data gaps, this could lead to false positives or false negatives. However, this would only affect the last data point before the gap and the first data point after the gap. For the IQR calculation, the method calculates the IQR over a time window. This means that it calculates the IQR from the available data points in the IQR window, and is therefore not affected by data gaps.

The edges of the time series would be the very first and the very last data point. The derivative function uses the difference of the first (last) two data points as the derivative at the beginning (end) of the time series, since it cannot calculate a central difference for these two points. Therefore, these points are still considered in the calculation of the derivatives, and hence also in the power law filter (Step 1A), and if the attributed difference exceeds the derivative limit in the derivative filter, they are flagged as contaminated. For the IQR filter method (Step 1B), we fill the first (last) data points with the calculated IQR value of the first (last) calculated data point. This means, the IQR is assumed constant for half of the IQR time window at the edges. In our case (with an IQR window of 24 h), this affects the first and the last 12 h of the dataset.

We realize that Eq.1 could lead to confusion. *dt* in Eq. 1 is the step size between two data points (see above). Since we are dealing with time series, this is equal to a time interval. However, mathematically, it is simply a step from one data point to the next. Therefore, *dt* does not have to be determined. We reformulated the description of the derivative filter in section 2.4.1 and added a paragraph where we discuss the edges and data gaps. Also, we decided to use the term "derivative" instead of the term "time derivative" and removed the time deltas. To be consistent, we renamed the gradient filter and call it now "derivative filter". We also included the description of the threshold filter in this paragraph, since the threshold filter is applied together with the derivative filter:

2.4.1 Step 1: Derivative filter

The derivative filter is used to separate periods characterized by rapid fluctuations in concentrations (we consider them as polluted periods), from those dominated by slow changes in concentration (we consider them as unaffected periods). At each data point in the native time series (10 s time resolution in our dataset) we calculate the absolute value of the derivative (i.e., change in concentration) of the concentration using the central differences formula.

$$|\boldsymbol{dC'_t}| \approx \left|\frac{\boldsymbol{C_{t+1}-C_{t-1}}}{\boldsymbol{2}}\right| \qquad \textbf{Eq. (1)}$$

where dC'_t refers to the derivative of concentration C at time t, C_{t+1} and C_{t-1} refer to the previous and following measured concentrations at time $(t+1)$ and $(t-1)$, respectively. Note that the derivative cannot be calculated with Eq. (1) at the edges of the dataset (very first and very last data points in the time series). Instead of the derivatives, the algorithm calculates the difference between the first (last) two data points at the beginning (end) of the dataset and uses those values for the derivative filter. This ensures that the edges of the dataset are also considered in the PDA. The derivative filter also ignores data gaps. For data points at the beginning and the end of a data gap, the derivative will still be calculated considering the previous and following data points, regardless of the duration of the gap (see Eq. 1).

To separate polluted from unaffected data we developed two methods:

Method A separates polluted from unaffected data with a power law. We average the derivatives of the particle number concentration over one minute (6 values) and plot them against the one minute-averaged particle number concentrations (Fig. 3). The averaging time can be varied and adapted to datasets with different time resolutions. This is discussed in Sect. 3.1. We choose one minute for a pragmatic reason: At one minute time resolution we can still see influences of short-lived changes in particle number concentration (e.g., from contamination) and it makes data processing faster as the size of the one-year long dataset is large. Figure 3a shows two "branches" of data points (visually emphasized by the relative wind direction color code): One with higher derivatives representing periods of high concentration variability, i.e., due to local contamination, and one with lower derivatives, indicating smooth variation, i.e., not affected by local contamination. Separating the contaminated and *unaffected branches is the fundamental step of the PDA developed here. The derivative of the particle number concentration can be described as a power law of the particle number concentration, and the two branches distribute around two different power laws. Thus, for the separation, we use a power law between those two branches*

$$(y = a * x^m) \qquad \text{Eq. (2)}$$

m corresponds to the slope, and $log(a)$ to the intercept with the logarithmic y-axis. Values for the power law fits are empirically selected.

Finding optimal values for a and m is an empirical process which can be validated by looking at the time series of the polluted and unaffected data together. This process likely needs several iterations until values for a and m are found which satisfy the needs of the intended data analysis. A higher slope in the separation line means that, for a fixed particle number concentration, the threshold of separation moves towards higher derivatives of particle number concentration, and therefore allows more variability in the data, i.e., the method is less strict. A higher intercept sets the threshold of separation to higher derivatives at lower concentrations, allowing for more variability there. Examples of four different separation lines are shown in Fig. 3a. For the MOSAiC dataset, we found a value of m = 0.55 s^{-1} and a = 0.5 $cm^{-3}s^{-1}$ (red line) to work well with our dataset (see Sect. 3.1).

*Method B separates data based on the moving centered interquartile range (IQR) of the derivatives within a defined time period (IQR window). Not all datasets show an equally clear separation of the derivatives into two branches like the particle number concentration shown in Fig. 3a. An example is the particle number concentration dataset from Jungfraujoch (Fig. 3b). An alternative method is thus to separate polluted from unaffected data based on the deviation of the derivatives from the centered IQR. For this, we calculate the centered IQR of the derivatives of each data point in a moving window of a set time period (IQR window) (24h in the case study described in Sect. 3.4.4, which is equal to 1440 data points). This means that for each data point, we calculate the IQR from the data +/- 1/2 of the IQR window before and after the data point. When the absolute derivative of a data point exceeds the 75th percentile by a given factor (hereafter called IQR factor), the data point is flagged. We use an IQR factor of 1.7 to identify contamination in the JFJ dataset. Both the IQR window size and the IQR factor can be adjusted in the PDA code. Method B is well suited to separate datasets with less obvious difference between contamination and unaffected periods. As a first start, we therefore suggest trying an IQR window size of 1440*x, where x is the time resolution of the dataset. We found the factor 1440 to work for datasets with 1 minute time resolution, where it represents a time window of 24 hours.*

Note that the moving centered IQR can only be calculated for data points with a distance of half of the IQR window from the edges in the dataset. To also account for the edges of the dataset, we fill the first (last) data points with the calculated IQR value of the first (last) calculated data point. This means that the IQR is assumed constant for half of the IQR time window at the edges. In our case (with an IQR window of 24 h), this affects the first and the last 12 h of the dataset.

Simultaneously with the derivative filter, we introduce an upper and lower concentration threshold, as described below, beyond which data are removed. Given the simplicity of this step, it is subsumed under the derivative filter step.

Here, we paste the paragraph about the threshold filters, which was formerly under section 2.4.2 and is now included in section 2.4.1.

R1.4

In the IQR method how is the duration of the moving window related to the dataset with respect to time resolution and expected duration of the anthropogenic events.

The window size of the IQR method depends on the time resolution of the dataset and on the spikiness of the signal and varies from dataset to dataset. We can therefore not come up with a strict number, but we can provide a hint for the user/reader, which is based on our experience with several different datasets. This is in analogy to the finding of the values of the derivative filter parameters. We added two sentences to section 2.4.1:

As a first start, we therefore suggest to try an IQR window size of 1440*x, where x is the time resolution of the dataset. We found the factor 1440 to work for datasets with 1 minute time resolution, where it represents a time window of 24 hours.

The whole new text of section 2.4.1 is presented in R1.3. As mentioned in section 4, we found the PDA to work well for datasets with a time resolution between 10 seconds and 10 minutes.

R1.5

In the IQR method a moving window is mentioned and hence a data point can be evaluated multiple times. It is not clear when it is flagged though. If it exceeds the IQR threshold once or multiple times. If it is the latter case how many exceedances should occur? Please clarify.

This is a good point, thank you for raising it. Every data point is only evaluated once, when it is in the middle of the moving IQR window. More precisely, for each data point we calculate the centered IQR in a given time window (IQR window) and compare the data point to this IQR. If it exceeds the IQR by a given factor (IQR factor), it is flagged. Therefore, every data point contributes multiple times to the IQR calculation, but it is only evaluated once. This has been clarified in the revised manuscript (see response R1.3).

R1.6

To further investigate the limitations of the PDA I am taking advantage that this is an open discussion and share two datasets to discuss how the algorithm behaves towards them. These are two case studies not discussed in the manuscript.

Dataset 1: It is frequent, eg due to A/C influence, that the standard deviation of the measurements changes abruptly even though the mean remains the same. How this case should be treated?

Dataset 2: It is assumed that any contamination would add to the local component. How does the algorithm treat data below the local component that are scarcely met but still exist?

Thank you very much for providing us with two datasets to test the PDA. We included two figures below to illustrate the behavior of the PDA on those datasets.

We assume that Dataset1 shows a contamination event between 11:30 and 13:30 and some polluted outliers before and after the event. It took us 4 iterations to come up with the figure presented below (Figure R1.6_1). The figure shows from a) to d) how the PDA flags data with the application of several filtering steps. For the first filtering step, we applied the IQR filter. This dataset is special because there are negative concentrations. To account for this, we set the lower threshold of the threshold filter to -2000. The applied parameters are shown in Table R1.6_1.

In Dataset2 it is not as obvious to differentiate between contaminated and unaffected data. It seems like the whole dataset is slightly affected by contamination. There is a spike just before 08:00, which we would want to flag, and also some higher concentrations around 10:00. We applied a stricter PDA to this dataset as you can see in Fig. R1.6_2 d). The parameters are shown in Table R1.6_1. However, it remains a discussion point whether one would rather discard the whole dataset. This depends also on the requirements of the user and on the context in which this dataset was collected.

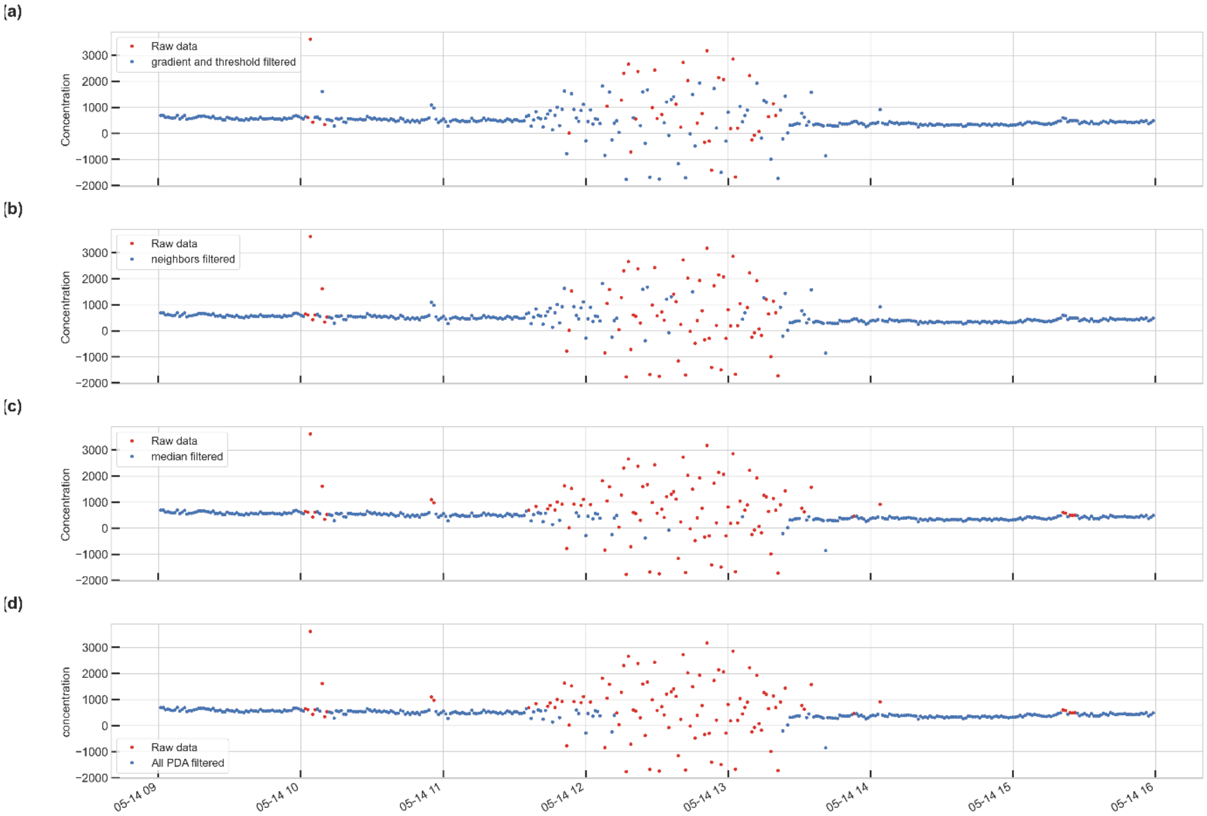

Figure R1.6_1: Dataset1, after the application of different filtering steps of the PDA.

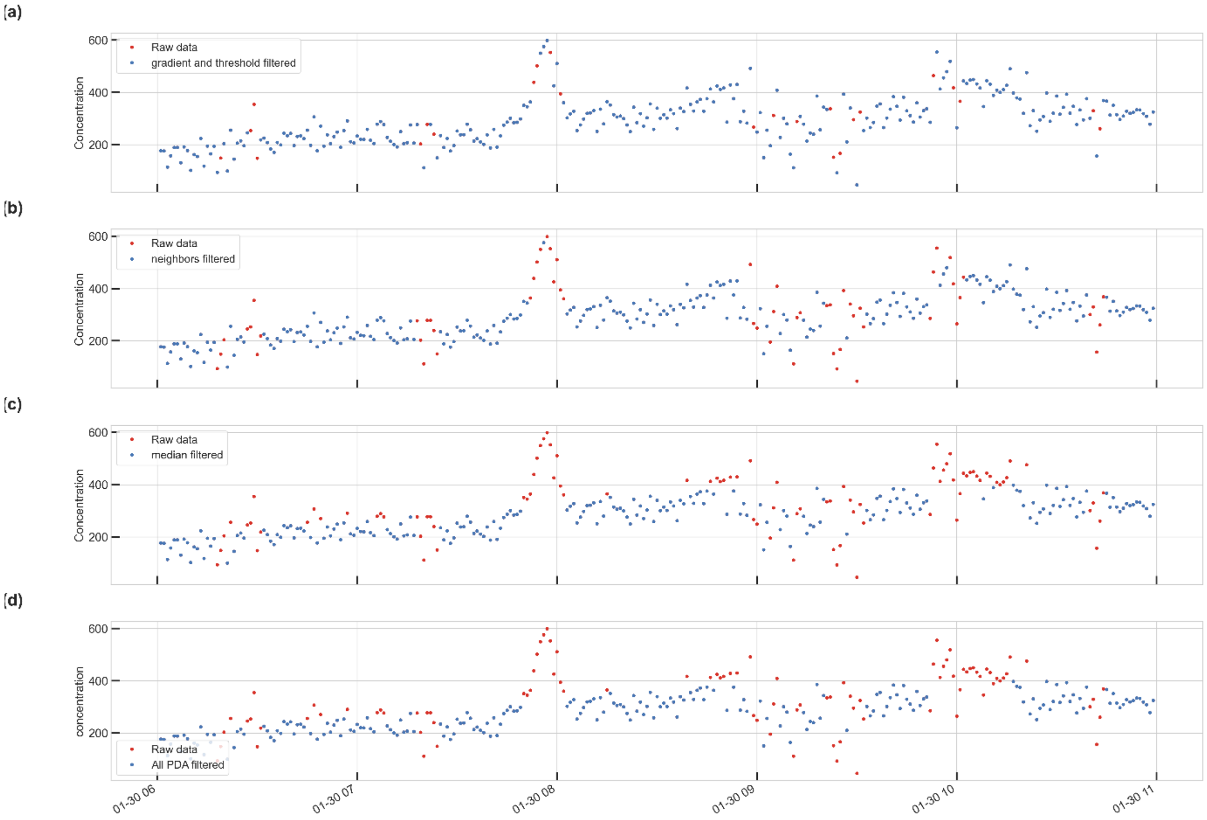

Figure R1.6_2: Dataset2 after the application of different filtering steps of the PDA.

Table R1.6: Summary of the parameters used to apply the PDA to the datasets. We left the units.

Parameter	Dataset1	Dataset2
IQR window	240 min	60 min
IQR factor	1.1	1.1
Upper threshold	1000	1000
Lower threshold	-2000	100
Neighboring points filter	Yes	Yes
Median time interval	120 min	120 min
Median deviation factor	1.3	1.2
Sparse window	30	30
Sparse threshold	24	20

R1.7

There are of course limitations not related to the algorithm but to the processes themselves. A major assumption is that the time resolution of the dataset should be higher than the duration of the anthropogenic influence.

This is an important point and it depends on the instrument’s sampling frequency and on the measurement type. Of course, if the anthropogenic influence has a shorter time than the instrument is able to detect, it would not “see” the influence. Often, the instrument’s sampling frequency is faster than the duration of the anthropogenic influences. For example, in our case the CPC dataset is representing average values over 10 seconds, even a shorter spike of 1 second would affect the averaged concentration.

R1.8

A flow chart with respect to the two data gradient methods should be added, to make clearer the algorithm process. Please use the standard schematics.

Thank you for pointing this out. We adapted the flow chart in Figure 2 and used the ISO 9001 flow chart scheme.

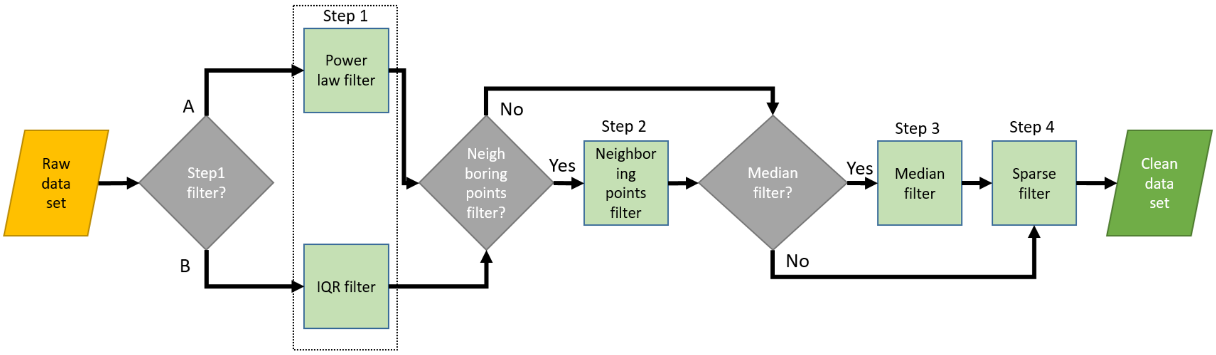

Figure 2: Schematic of the pollution detection algorithm. The key is the power law filter (highlighted in a dotted rectangle), which is followed by a series of steps. The neighboring points and the median filter are optional and can be skipped. Parameters of each step can be adjusted. IQR stands for interquartile range (see Sect. 2.4.1).

R1.9

As discussed in the manuscript, there are difference of the PDA and the manual method, which relate to false positive (measurements not identified as polluted even though they are) and false negative (non polluted measurements identified as polluted). Please include in either Table 1 or 2 how many false negatives and positive, compared to the visual method, the PDA leaves behind in each step.

Unfortunately, we cannot compare these numbers in table 2, since this describes the dataset of the CPC3025, which was not cleaned visually. Table 1 is dedicated to sum all factors and thresholds in each step for different datasets. However, Table 3 already shows a comparison between the visual method and the PDA, based on the data of the CPCf. Please note that only the CPCf was cleaned visually. We added the number of false positives (PDA detects contamination, but visual cleaning does not) and false negatives (PDA does not detect contamination, but visual cleaning does) of the PDA compared to the visual method. Note that the numbers for PDA clean (+ 5 data points) and both clean (+ 30 data points) slightly changed due to a typo in the calculation of the numbers. However, these numbers are so small that they did not affect the percentages of the two contributions.

Table 3: Fraction of clean data points of the derivative filtering method and the visual filtering method compared to the total number of data points (total counts) in numbers and in percent of the total counts. This table is based on the CPCf dataset in 1 min time resolution.

	# Data points	Percentage

Total counts	308750	100.00%
PDA clean	197671	64.02%
PDA polluted	111079	35.98%
Visual inspection clean	214540	69.49%
Visual inspection polluted	94210	30.51%
PDA clean, visual polluted	947	0.31%
PDA polluted, visual clean	17816	5.77%
Both clean	196724	63.72%
Both polluted	93263	30.21%

R1.10

How are the false negatives and positive distributed? Is there a pattern or are they random?

As shown in Fig. 9, false negatives and false positives are distributed over all concentrations. The PDA detects slightly more contamination than the visual cleaning method (5.77%), it is stricter. Note that in figure 9 the red contour line slightly changed, because we found a typo in the calculation of the number of data points in the comparison (see R1.9).

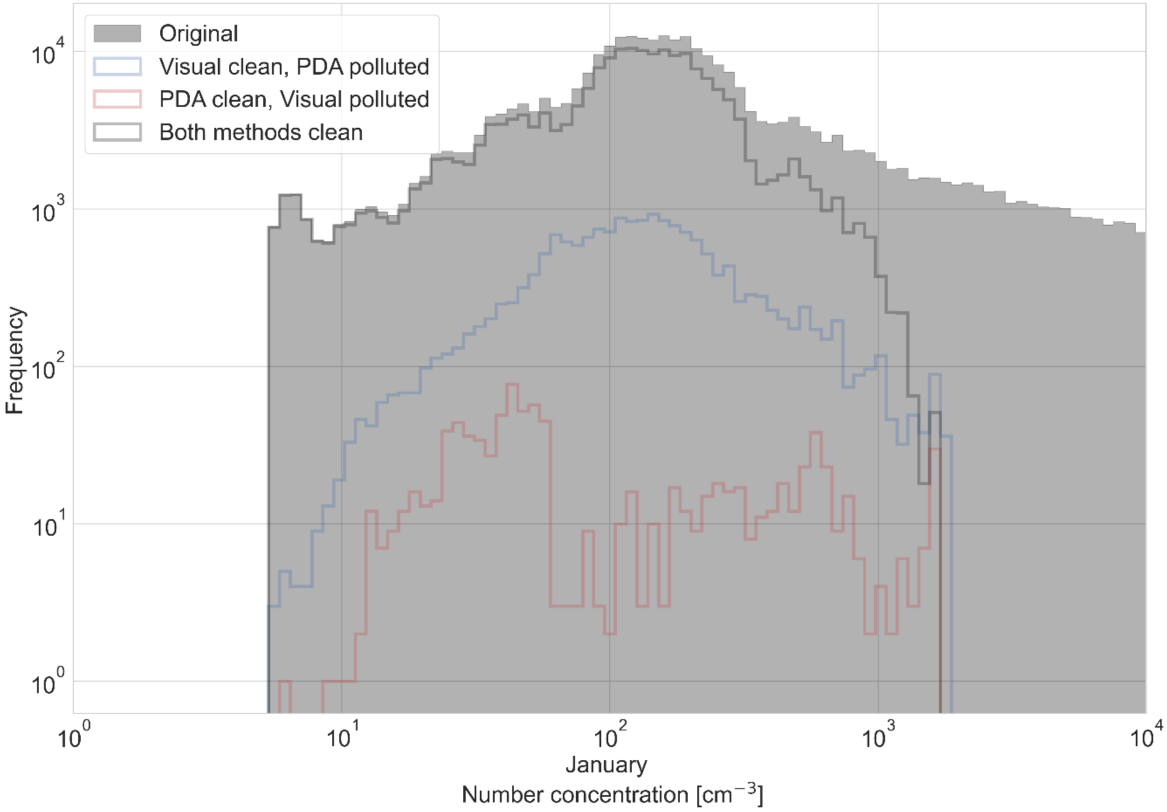

Figure 9: Comparison of the visual inspection method to the PDA on the dataset of the CPCf of ARM. Original data are shown in grey. The blue contour line shows the fraction of data points where only the visual inspection method, but not the PDA, considered data to be clean (6 %). The red contour line shows the opposite, i.e., the fraction of data points where only the PDA, but not the visual inspection method, considered data to be clean (<1 %). The dark grey contour line shows the fraction of data points where both methods considered data to be clean (~64 %).

R1.11

A discussion on how varying each of the 7 parameters affects the amount of false negatives in a case study presented in the manuscript would be beneficial.

This is a very good idea. We revised section 3.1 and compare the effect of individual filtering steps in the PDA and how changes of parameters in different steps affect the number of unaffected data points. There we show that the derivative filter flags the largest part of the data and that the other filtering steps have a fine tuning effect on the PDA. The finetuning effect can be important to detect false negatives from the derivative filter. We show the complete revised section 3.1 in R 2.4.

R1.12

A method quite similar to this work has already been published (Gallo et al., 2020). Also additional methods have been applied, such as smoothing, to mask short term local events (Liu et al., 2018). This is a subject the community has spent some time to investigate and there is some literature out there. Most notably Giostra et al., 2011; McNabola et al., 2011; Brantley et al., 2014.

Brantley, H. L., Hagler, G. S. W., Kimbrough, E. S., Williams, R. W., Mukerjee, S., and Neas, L. M.: Mobile air monitoring dataprocessing strategies and effects on spatial air pollution trends, Atmos. Meas. Tech., 7, 2169–2183, https://doi.org/10.5194/amt-7-2169-2014, 2014.

Gallo, F., Uin, J., Springston, S., Wang, J., Zheng, G., Kuang, C., Wood, R., Azevedo, E. B., McComiskey, A., Mei, F., Theisen, A., Kyrouac, J., and Aiken, A. C.: Identifying a regional aerosol baseline in the eastern North Atlantic using collocated measurements and a mathematical algorithm to mask high-submicron-number-concentration aerosol events, Atmos. Chem. Phys., 20, 7553–7573, https://doi.org/10.5194/acp-20-7553-2020, 2020.

Giostra, U., Furlani, F., Arduini, J., Cava, D., Manning, A. J., O'Doherty, S. J., Reimann, S., and Maione, M.: The determination of a "regional" atmospheric background mixing ratio for anthropogenic greenhouse gases: A comparison of two independent methods, Atmos. Environ., 45, 7396–7405, https://doi.org/10.1016/j.atmosenv.2011.06.076, 2011

Liu, J., Dedrick, J., Russell, L. M., Senum, G. I., Uin, J., Kuang, C., Springston, S. R., Leaitch, W. R., Aiken, A. C., and Lubin, D.: High summertime aerosol organic functional group concentrations from marine and seabird sources at Ross Island, Antarctica, during AWARE, Atmos. Chem. Phys., 18, 8571– 8587, https://doi.org/10.5194/acp-18-8571-2018, 2018

McNabola, A., McCreddin, A., Gill, L. W., and Broderick, B. M.: Analysis of the relationship between urban background air pollution concentrations and the personal exposure of office workers in Dublin, Ireland, using baseline separation techniques, Atmos. Pollut. Res., 2, 80–88, https://doi.org/10.5094/APR.2011.010, 2011

Thank you for pointing this out. We have included reference to the previous work in the introduction. This will be inserted in line 69, after the introduction of the REBS method:

Liu et al. (2018)*, used a de-spike algorithm, based on a 24 h running median window, to remove short-term local contamination events of less than 1h duration from an aerosol time series measured at*

McMurdo Station in Antarctica. Giostra et al. (2011) *used a statistical approach where they extract the baseline with a decomposition of the probability density function of the data. Polluted data shows a gamma distribution, the baseline is represented as a Gaussian distribution. This method was applied on halocarbon data from remote marine or alpine stations. El Yazidi et al. (2018) applied the REBS method to four datasets of trace gas measurements and compared it to the standard deviation (SD) method for particles (Drewnick et al., 2012), which detects contamination as data points that differ by more than 3σ from the median of the data, and to the coefficient of variation (COV) method (Hagler et al., 2012), which uses the 99th percentile of the COV as a threshold for contamination. Hereby, the COV is defined as the standard deviation of a moving time window (5 min), divided by the mean value of the whole dataset.* Brantley et al. (2014) *compared the SD method to the COV method to detect exhaust plumes from air quality measurements on a road. Both these methods work for datasets in which the signal of plumes is characterized by high variability and magnitude (Brantley et al., 2014).* McNabola et al. (2011) *applied baseflow separation techniques, such as low pass filters, or moving interval filters, known from stream-flow hydrology, to separate background concentrations in urban PM10 measurements and compared the result to background PM10 measurements. Gallo et al., (2020) developed a method to retrieve the regional aerosol number concentration baseline at the Eastern North Atlantic (ENA) Atmospheric Radiation Measurement (ARM) user facility from the U.S. Department of Energy's (DOE). The ENA Aerosol Mask (ENA-AM) identifies data points, which exceed the standard deviation of the data below the median (σ_b) of a 1- month period by more than a factor α. They found the method to work best for time periods between two weeks and one month, and less than half of the data points influenced by local contamination.*

Reviewer2

Review of "Automated identification of local contamination in remote atmospheric composition time series," by I. Beck et al., submitted to Atmospheric Measurement Technology.

The manuscript describes a Pollution Detection Algorithm (PDA) that consists of a number of filters that can applied to a time series of measurements to eliminate those values that are influenced by pollution without use of ancillary data such as CO concentration that might assist in this effort. The algorithm consists of 5 sequential sets of filters. The first, and primary one, is referred to as the gradient filter (although gradient typically refers to a spatial derivative, so I would recommend a better name for this), which removes points for which the time derivative of a concentration is greater than a given value that might depend on the concentration itself. The next filter is the threshold filter, a simple cutoff above which all data are classified as polluted. The others are a neighboring points filter that removes points at the start or end of ones that are flagged as polluted, a median filter that removes points that exceed the running median by a given factor, and a sparse data filter that removes points that are surrounded by ones that were removed. Several examples of time series to which the PDA was applied were presented.

I cannot recommend publication of the manuscript as it stands, as the arguments that the PDA is especially novel or necessary were not compelling. I provide some suggestions below that would if followed would allow me to reconsider this decision. Basically, the manuscript should be a bit more explicit in what it is trying to do and there should be comparisons among different approaches so that the utility of the PDA can be evaluated and demonstrated. I will provide some general comments followed by a number of more minor items.

We thank the reviewer for their critical remarks, which indeed helped us to greatly improve the manuscript.

The purpose of the PDA is to provide an automated tool which avoids manual cleaning as well as a versatile pollution recognition method that exceeds the results of single filter methods such as e.g. a median filter. The former point is key for large datasets from campaigns and permanent monitoring stations, where manual cleaning is often beyond the capacities of the data originators. The latter point is explained in more detail in answer R2.4, which compares the PDA with a number of other approaches, i.e. a simple threshold filter, only a median filter, only a derivative filter, and fast Fourier Transform (new appendix B). At the end of the introduction, we now phrase more explicitly the purpose of the PDA.

We have also replaced the word "gradient" by "derivative".

We add a copy of the paragraph which we included in the introduction:

This makes the algorithm efficient and consistent way to detect local contamination in large remote atmospheric time series, as they exist for example from ship campaigns or from remote stations. This method is objective as the treatment of the data is consistent throughout the whole time series considered, because the same value of each parameter is applied to the entire dataset.

GENERAL COMMENTS

R2.1

The authors assume throughout that in remote regions there is a background signal that is slowly varying and that any pollution will manifest itself by the presence of spikes. They refer to this "background" as a "baseline," this terminology is a poor choice, and they should not switch between the two (and should not use "baseline" at all).

Thank you for these remarks. We changed the terminology to describe the well-mixed background signal to the term "background" and avoid the use of "baseline" in the entire manuscript. We added the following sentences to the very beginning of section 2 to clarify what we mean when we talk about "contamination" and "background":

In this manuscript, we use the terms "contamination" and "pollution" interchangeably to describe local contamination. We define local contamination as fresh exhaust plumes from the ship, skidoos, snow groomers and other local, anthropogenic sources of pollution. We define the background concentration as unaffected from local contamination and well-mixed ambient concentrations. This means that background observations can contain aged pollution, e.g. an aged plume which is long-range transported to Polarstern (Dada et al., 2022). Note, that the aim of the PDA is to identify fresh local contamination and we do not aim at detecting aged, well-mixed contamination.

R2.2

Although they start by discussion pollution in remote environments, this reviewer was left with the impression that the approach was designed to be used more universally; for instance, on line 101 they state "a common filtering method, which relies on a minimal number of input variables, is desirable to achieve reproducible pollution detection across a variety of datasets," and on line 106 they state "the method can be applied to a large number of measurement sites." If this approach is restricted to remote environments where local contamination occurs only in the form of spikes and higher frequency signals, then what is presented is essentially a spike-removal, or smoothing routine. My immediate thought was why not do an FFT, remove the high-frequency components, and revert the data to a smoothed time series.

Thank you for pointing out that the purpose of the PDA was not clearly enough presented. The motivation for developing the PDA is indeed measurements in remote environments, where local contamination often manifests itself as spikes and periods of very high concentrations, compared to the background. We found, however, that the PDA can be applied to other datasets as well, as demonstrated for the Jungfraujoch, which makes it a more versatile tool and can serve a larger community.

Local contamination can manifest itself as single spikes with strong concentration changes or as longer time periods of up to several days with strongly fluctuating concentrations (see Fig. A8) as well as noise (see Fig. A11). These periods can be interrupted by short periods of unaffected concentrations, which we aim to retain. Therefore, there is no single cut-off frequency, which could be used for a low pass filter. This is shown in the derivative vs total concentration figure (Figure 3): The derivative value increases with the total particle number concentration and this is the reason why we cannot use a fixed derivative threshold but we adjust it following a power law. Below we attached a periodogram of the measurements of the CPC3025 in March from MOSAiC, where it is apparent that we cannot apply a low pass filter. For the above mentioned reasons we also cannot simply apply a spike detection algorithm.

We would like to emphasize that the PDA is not a smoothing routine, but an algorithm to detect local contamination in time series.

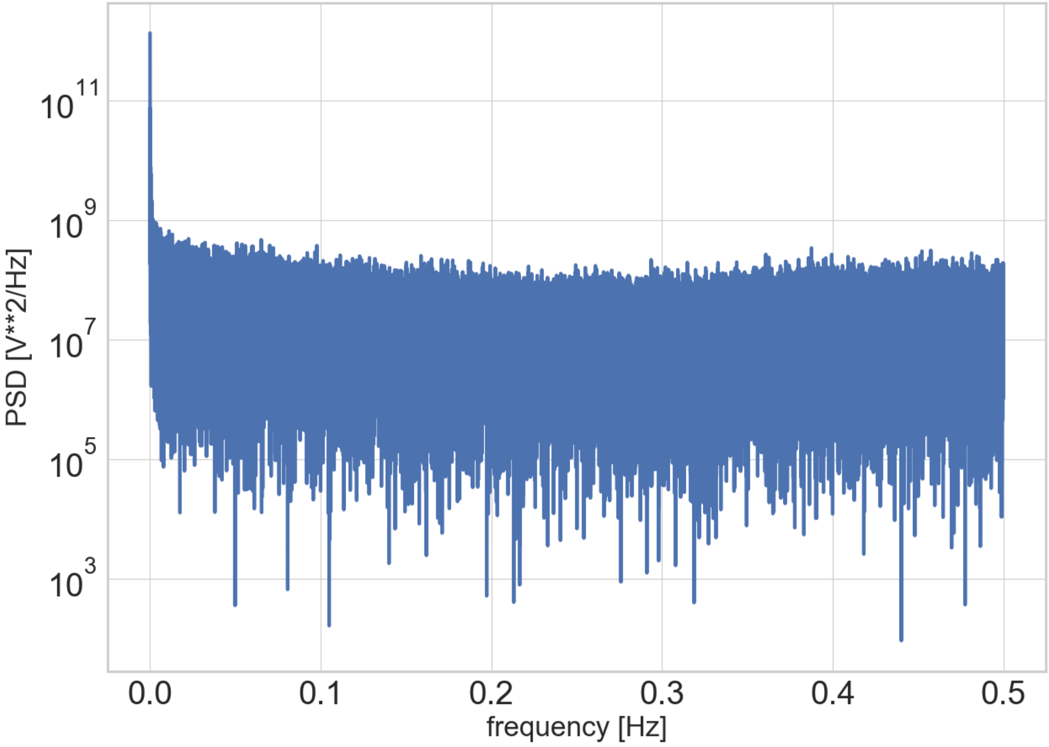

Figure R2.2: Power spectral density (PSD) of the particle number concentrations of the CPC3025 as a function of the frequencies. The dataset has a time resolution of 10 seconds. For this figure we used the subset of the month March from MOSAiC.

R2.3

The PDA does seem to work in that it removes a large number of data points that visual inspection would also remove, and the examples presented demonstrated that for a shipboard deployment the PDA was better than selecting only by wind direction. However, the manuscript did not make a compelling case that the PDA is necessary or that it is superior to visual inspection, a simple threshold approach (the second of their five sets of filters), or a median-type approach (the third of their five sets of filters). The argument was made that the PDA would be easier to apply and less subjective than a visual approach, but the number of adjustable parameters (I counted 8) that require specification and selection among various options argue against the method being a totally objective one, and it is still necessary to examine the results to ensure they look reasonable. This is noted on line 219, where it is stated: "Every pollution filtering method contains a certain level of subjectivity since the final decision about polluted vs non-polluted must be made by the user."

The reviewer raises good points, most importantly the direct comparison to more standard methods such as a threshold filter only, a median filter only, or others to motivate more clearly the necessity to develop a new algorithm. Please see answer R2.4 for details on the comparison between methods.

The purpose of the PDA is to achieve similar or near-equal results as the visual inspection with much less effort. Here we take the visual inspection as the reference. As stated in section 2.3, line 219, the final decision if a data point is contaminated or not, must always be made by the user, since the true background is not known. The advantage of the PDA, compared to the visual inspection is that it is not biased because it applies the same criteria to the whole dataset, hence is a fully consistent and reproducible approach. When cleaning visually, the criteria can change from point to point without the intention to do so. Removing contaminated data points visually can also be very time consuming with larger datasets and it cannot be reproduced by other scientists, nor can it be applied simultaneously to multiple datasets for comparison reasons.

Regarding the point of objectivity, we repeat here the answer R1.1 to reviewer 1 who had a very similar comment: The reviewer is correct that setting between 3 and 7 parameters is not entirely objective. However, the treatment of the dataset in itself is consistent, because these parameters are not changed. All parameters are set once in the beginning and applied to the whole dataset. Hence the algorithm is more time efficient and more consistent in itself in terms of strictness, compared to, for instance, a manual cleaning method, where the expert has to decide on each individual data point. The 3 to 7 parameters also make the algorithm flexible. We aim to provide an algorithm, which can be used with many different remote atmospheric datasets. This comes with the trade-off that the user needs to adjust all parameters. In other words: If we restrict the parameters, we exclude more datasets and more use cases, hence the algorithm could not be applied to many other user's datasets. To emphasize the advantages of the PDA, and to clarify what is meant by "objective" we have made a change in the manuscript in Section 1, which we copy here (see also R2.1):

This makes the algorithm an efficient and consistent way to detect local contamination in large remote atmospheric time series, as they exist for example from ship campaigns or from remote stations. This method is objective as the treatment of the data is consistent throughout the whole time series considered, because the same value of each parameter is applied to the entire dataset.

R2.4

The gradient approach did not work in some of the examples presented, and in those a simple threshold filter seems as though it would work quite well (especially for the CO2 time series). The threshold filter was their second option, but that is not sufficiently innovative by itself to justify publication. Likewise, the manuscript did not demonstrate that an approach similar to what they termed the "median filter" such as a simple filtering method that removed points greater than, say, 2-sigma about a moving average would not have performed as well and given the same results as their PDA, and in the examples presented, it seemed as if it would.

I would have preferred to see comparisons made among 1) the gradient filter, 2) visual inspection, 3) a simple threshold, and 4) a median, or 2-sigma moving average for one or more given data sets (better yet, selected time series where the comparisons can be meaningfully evaluated), and a discussion of which is better and why. The comparisons that were presented did not allow evaluation *of the utility of the gradient method, which is their first filter. Why not a median filter first, for instance? It seems as though it would do just as well.*

This is a very good point, and we included a comparison of individual filtering steps in the appendix. For the sake of clarity, we decided to only compare single filtering steps of the PDA, in order to demonstrate why we chose to include all steps in the PDA and single filtering steps would not work to our satisfaction. This is why a new method, the PDA is needed, which is a novel tool. We also added a paragraph to discuss the application of a low pass filter to the data, together with the figure of the spectral density of the frequencies.

Section 3.2.2 discusses the intercomparison of the PDA to the visual cleaning method. As mentioned in R2.3, we do not aim to provide a better filtering method than the visual filtering, but rather use the visual filtering method to compare the results of the less time-consuming PDA to. This was discussed in R1.9, and we adapted Table 3 to provide a better overview of the differences between the PDA and the visual filtering method (see R1.9).

Indeed, we applied all filtering steps of the PDA as stand-alone methods to our time series, and the derivative filter shows by far the best performance. We deal with a one-year long dataset. Within this year, the background concentration changes from <10 cm^{-3} in winter to up to 5000 cm^{-3} in summer. This signal is often disturbed by locally produced contamination, mostly from the ship's stack, but also from flying helicopters or engines running on the ice around the ship. This makes the application of a simple threshold impossible for the whole dataset. The same applies for the CO_2 time-series; the strong seasonal cycle (419.9 $\pm$ 1.5 ppm in January (mean $\pm$standard deviation) vs. 401.7 $\pm$ 1.1 ppm in August) makes it impossible to apply a simple threshold for the whole dataset.
In highly contaminated situations, the derivative filter leads to false negatives, because the CPC only reports its maximum or near-maximum values, hence there is little variation in the concentration. By defining an upper threshold, we flag those data points already at an early stage of the PDA process. The lower threshold, in contrast, prevents the derivative filter from flagging data points with very low particle number concentrations which are located in between highly contaminated data points or related to precipitation events during unaffected periods. This helps to retrieve a background signal even during polluted periods.

The median filter is good at finding outliers in a time window, which is dominated by the background concentration. Our data are often dominated by local contamination over longer time periods, which makes the application of a median filter alone impossible. Therefore, the median filter is only effective when applied to a pre-cleaned dataset. This is also the reason why we apply the median filter after the derivative filter.

The order of the filtering steps follows the logic: First, we apply the most powerful filter (the derivative filter). Then we use the following filters for the "fine tuning", to remove false negatives (data points which are contaminated but were not detected by the derivative filter). As mentioned above, the median filter would not be effective if applied to the raw dataset.

We added the following text to section 3.1:

To assess the effect of each filtering step, we applied each of them individually to the CPC3025 dataset and discuss this in Appendix B:

Appendix B: Comparison of individual filtering steps

In figure B1, we compare how the application of each individual filtering step to the 1min resolution dataset of the CPC3025 performs on the case study from March 6th to March 8th. Panel a) shows the *result after the application of the derivative filter and the lower threshold filter only (but not the upper threshold filter) with a = 0.5 and m = 0.55 and a lower threshold of 60 cm^{-3}. As we can see, the application of the derivative filter detects and flags most data points during the polluted time periods, but leaves some during the contamination event on the 6th of March. The application of the derivative filter leaves 43% of the data unaffected and it reduces the mean concentration from 5198 cm^{-3} to 202 cm^{-3}. Panel b) shows the application of the upper threshold filter alone. Here we set the upper threshold to as low as 130 cm^{-3} to be able to retrieve the background signal as much as possible. With this threshold, 23% of the data are left unaffected with a mean concentration of 70 cm^{-3}. However, the application of a single threshold to a longer time series is difficult, since the background concentration can rise to higher concentrations (as can be seen for example in Fig. 6). The upper threshold can be useful in cases, when the measured concentration stays at the upper detection limit of the instrument over a long time period and thus the derivative filter would not catch those contaminated data points. Panel c) shows the application of the median filter alone with a median window of 360 data points (6 hours) and a median threshold of 1.05. The application of the median filter alone with these parameters leaves 68% of the data unaffected, with a mean concentration of 2979 cm^{-3}. It is not satisfying because it is not able to flag the strong contamination on the 8th of March after 12:00. Too many contaminated data points raise the median concentration. The median filter relies on a pre-cleaned dataset, where most of the contaminated data points have been removed already. Therefore, it can only be applied after the application of the derivative filter. Finally, Panel d) shows the result after the application of the whole PDA, with the parameters presented in Table 1. The application of the whole PDA leaves 38% of the data unaffected with a mean concentration of 191 cm^{-3}. Evaluated visually by expert's judgement, we find that it performs better than the application of the single filters, it detects more contaminated data points and results in a time series which represents the background concentration. Table B1 shows an overview of how many data remain unaffected after the application of the different filtering steps. Additionally, the mean concentrations and the standard deviations are shown. The derivative filter is by far the most powerful step of the PDA, as it detects already 64% of the total contamination and reduces the mean concentration drastically. The other filters of the PDA only have a "fine-tuning" effect and add another 6% of flagged data points. This effect can still be very important for individual cases as shown in the case study during March 6 around noon (Fig. 4).*

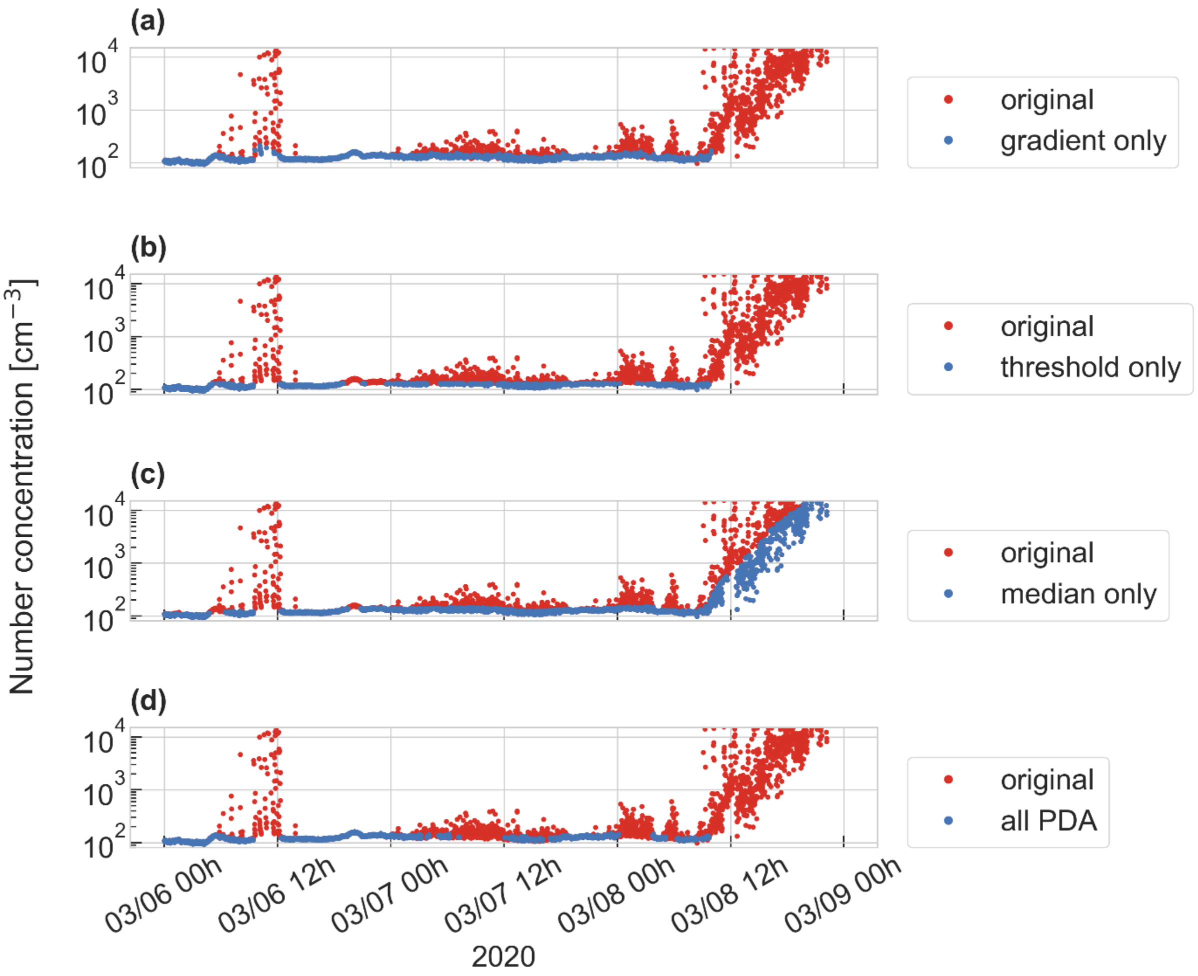

Figure B1: Intercomparison of individual filtering steps on a case study of March 6 to 8. Clean data (in blue) is overlaid over the original data (in Red) after the application of one filtering step individually to the data: a: derivative filter. b: threshold filter. c: median filter. d: All filtering steps of the PDA were applied. For all plots we used data from the CPC3025 at 1 min time resolution. Original data have only been pre-cleaned for zero filter measurements.

Table B1: Percentage of data declared as unaffected when different filtering steps are applied and the mean concentrations and standard deviations of the corresponding particle number concentrations.

Comparison of single filters	**Parameters**	**Remaining data**	**Mean concentration [cm^{-3}]**	**Sdandard deviation**
Total counts		100%	5198	14598
Derivative filter only	a = 0.5 , m= 0.55	43%	202	618
Threshold filter only	Threshold = 130 cm^{-3}	23%	70	37
Median filter only	Median time = 360 min , median factor = 1.05	68%	2979	10646
Derivative and threshold filter	As in Tab. 1	43%	198	244
Derivative, threshold and neighbors filter	As in Tab. 1	39%	191	221
All PDA	As in Tab. 1	38%	191	214

Since local contamination often shows in fast changing concentration spikes, it is worth exploring of whether a low-pass filter is applicable. For this, we looked at the power spectral density of the CPC3025 particle concentration data by means of a Fourier frequency decomposition (Fig.B2). No high frequency is visible which would allow a low-pass filter to be applied. Local contamination in this dataset does not show in a high-frequency signal, which is distinguishable from the background signal. The detection of pollution based on frequency analysis is therefore not possible.

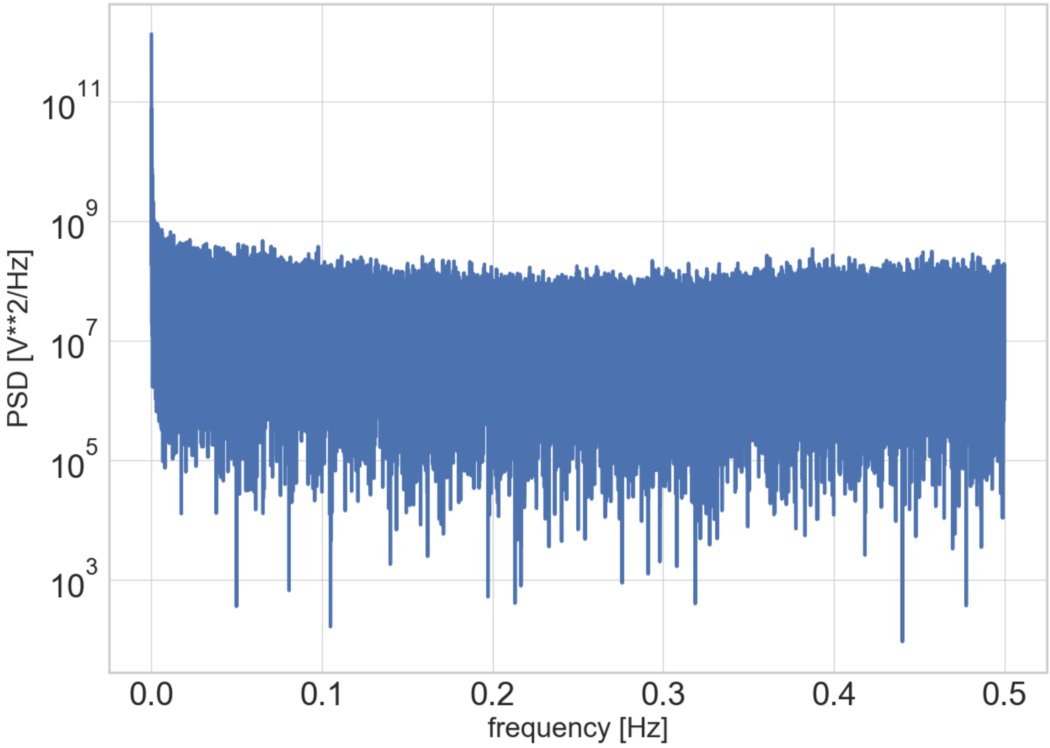

Figure B2: Power spectral density (PSD) of the particle number concentrations of the CPC3025 as a function of the frequencies. The dataset has a time resolution of 10 seconds. For this figure we used the subset of the month March.

In order to elaborate on the effect of changes in the parameters of individual filtering steps, we let the PDA run several times and thereby only change one parameter at the time. The resulting size of the filtered dataset is shown in Table B2. The first row shows the initial setting, as we used them in Table 1. For example, the largest change is caused by turning off the neighbors filter. This increases the dataset by 11.4%. Relatively small changes in the power law slope and intercept of the derivative filter change the size of the dataset by roughly 5-10 %, whereby the effect of changes of the slope are stronger. Changes in the median filter only cause small changes by < 1% to the final dataset. And setting the sparse threshold from 24 to 18 out of 30 data points (from 80% to 60% allowed polluted data points in the sparse window) reduces the dataset by ca. 3%. The table illustrates again that the derivative filter is responsible for the largest part of the filtering by the PDA. Even though the filtering steps 2 to 4 only contribute little to the PDA, they are valuable to avoid false negatives after the application of the derivative filter.

Table B2: The effect of changes in the parameters of individual filtering steps on the number of unaffected data points. The first row shows the standard settings used to filter the CPC3025 dataset and the number of remaining data points. The following rows show changes in different parameters and again the number of unaffected data points with these changes in the PDA.

Case number	**Initial parameters of the PDA**	**# Data points after application of PDA**	**Percentage**
A	a= 0.5 cm^{-3}s^{-1} m= 0.55 s^{-1} lower threshold = 60 cm^{-3} median time interval = 30 min median deviation factor = 1.4 sparse window = 30 sparse threshold = 24	190358	100.0%
	Changed parameter		
	a= 0.45 cm^{-3}s^{-1}	184297	96.8%
B	a = 0.6 cm^{-3}s^{-1}	198733	104.4%
	m = 0.5 s^{-1}	171060	89.9%
	m = 0.6 s^{-1}	202292	106.3%
	lower threshold = 100 cm^{-3}	196471	103.2%
C	median time interval = 120 min	188503	99.0%
	median_factor = 1.8	191316	100.5%
	median_factor = 5	191893	100.8%
	sparse_threshold = 18	185578	97.5%
D	sparse_threshold = 27	192761	101.3%
	no neighbors filter	212073	111.4%
	no sparse filter	193680	101.7%

R2.5

The authors should be explicit in what they mean by pollution and local sources, as there is not a clear demarcation between these, and the definitions used were often operational; for instance, pollution being manifested by large spikes. On line 252 it is stated "Generally, concentration data from remote regions, characterized by the absence of dominant local (anthropogenic) sources, vary only slowly with time." Similarly, on line 255 they state "The PDA builds on this abrupt variation in concentration and detects polluted data based on the rate and magnitude of change in the concentration signal over a given time period," which is the crux of their method. However, I can imagine a situation where *their ship moved into a day-old exhaust plume of another (or the same) ship that was well mixed and thus resulted in smooth concentrations without spikes. This would be a polluted situation, but is it local?*

If the data were not removed by a gradient filter, then the remaining points would not accurately represent the "background." The authors discuss this obliquely on line 85, where they mention recirculation of emitted pollution, and on line 249, where it is stated "Pollution influence can also occasionally be so small that it would not surpass the threshold," but a clearer expression of what their PDA can do should be stated.

Thank you for this comment, which is very valuable and touches an important question: at what point does contamination no longer count as local contamination? We aim at detecting fresh, locally produced plumes, created from the measurement campaign itself. We do not aim at detecting aged and well-mixed pollution, which influences the background concentration as well. To clarify what we mean with pollution and contamination, we copy our answer to R2.1 with the definition of these terms in section 2:

In this manuscript, we use the terms "contamination" and "pollution" interchangeably to describe local contamination. We define local contamination as fresh exhaust plumes from the ship, skidoos, snow groomers and other local, anthropogenic sources of pollution. We define the background concentration as unaffected from local contamination and well-mixed ambient concentrations. This means that background observations can contain aged pollution, instance e.g. an aged plume which is long-range transported to Polarstern (Dada et al., 2022). Note, that the aim of the PDA is to identify fresh local contamination and we do not aim at detecting aged, well-mixed contamination.

Our statement in line 249 refers to the threshold filter. What we mean and what we explain in more detail in R2.4, is that the threshold filter would not be able to detect local contamination in all situations.

R2.6

More fundamentally, the method was not really validated. This would be very difficult to do, as it would require some a priori knowledge of what is a polluted signal and what is not, but the manuscript seemed to imply that the result from their PDA is the background (i.e., non-polluted) result and using that as the gold standard against which to compare other methods. Better, in this reviewer's opinion, would be a comparison among the four approaches (visual, gradient, threshold, and median), as noted above.

We have followed the reviewer's suggestion and compared individual filtering steps (derivative, threshold and median) to each other in the appendix and the whole PDA to the visual filtering method in Section 3.2.2. Please see the detailed answer in R2.4.

MINOR COMMENTS

R2.7

It is not clear that the title is the most appropriate one. It states "local contamination" but refers throughout to "pollution", which might not necessarily be local.

We addressed it by removing the term "pollution" in many cases and replaced it with the term "contamination" or "local contamination". We describe our understanding of "local contamination" at the beginning of section 2, see also R2.1 and R2.5. However, in some cases, we decided to keep the word "pollution", where we think it is adequate.

R2.8

Line 103: By not including ancillary data sets (such as BC concentrations), the method is basically a spike-removal algorithm. The manuscript is attempting to sell the PDA as a one-size-fits-all approach, but in most cases, more information (e.g. inclusion of ancillary data sets) is better than less information.

We agree that more information is better than less. In answers 2.2 and 2.4 we show that the PDA goes beyond a simple despike method, because several features are necessary to clean the data satisfactorily. However, often there is no option to deploy a large number of instruments, particularly not in remote environments for a long time period. Hence our objective was to devise a method which works well based only on concentration data from instruments that are sensitive to local contamination such as a CPC or CO2 sensor. Ancillary datasets can always be used to verify the quality of the PDA. We would like to highlight that CO for example is not a good tracer for ship exhaust (see e.g., Figure A4), while it might be for other combustion sources.

R2.9

Line 260: this was said earlier (near line 103)

We removed it.

R2.10

Line 263: stated earlier on line 105

We removed it.

R2.11

Line 235: The text abruptly switches between discussions of the contamination sources of the data sets evaluated to a description of the algorithm and its availability to users.

The reviewer likely refers to line 264. We have removed this sentence from here and added it to the introduction towards the end.

R2.12

Line 265: This is where Section 2.4 should start, not after a discussion of data sets.

Thank you for pointing this out. We have now moved the first part of 2.4 (before line 265) to *2.1.2 Description of particle number concentration characteristics*. That means the subsections after 2.1.2 were shifted.

R2.13

Line 269: There is no section 2.3.4; this should be 2.4.4.

Thank you for the hint, this was fixed.

R2.14

Line 298: Averaging 10-s data over one minute yields an average of 6 values.

This is true, we specified it in the text and added the number of values over which we average. We prefer to speak of time-averages when dealing with these large time series, therefore we average over time and not over a number of points.

R2.15

Line 301: As data are taken at 10 sec, averaging over a minute reduces the time by a factor of 6, but is this really a concern for computational speed?

Well, yes, it does decrease the dataset, and for a one-year long dataset with >2.8 Million data points it reduces its size significantly.

R2.16

Line 309: Values for the power law fits are empirically selected.

We adapted this.

R2.17

Line 316: It was noted a few lines earlier that the fit was empirical.

We deleted the sentence on line 309 to avoid a repetition, thank you for showing it. Reviewer 1 also had some suggestions for section 2.4.1. We show the revised section in the answer to R1.3.

R2.18

Line 310: Is this explaining how to find coefficients of a power law fit from two points? I assume most readers know how to do this, so I would recommend leaving it out.

We followed your advice, the revised section 2.4.1 can be seen in the answer to R1.3.

R2.19

Line 316: Presumably the authors mean "validated" rather than verified (there is no sense in verifying that the fit is empirical), but to do so by looking at the time series implies that polluted data can be removed by eye.

This is correct, we replaced the word verified by validated. Yes, polluted data could be removed by eye (as the visual filtering method describes in section 2.3). Also, to validate the functionality of the PDA, the user needs to look at the result visually and decide about the quality of the detection. Please refer to R2.3 for a detailed answer on this topic.

R2.20

Line 375: Presumably the authors mean the "number of polluted data points"; the "sum" is ambiguous and potentially confusing.

Thank you, we followed your recommendations.

R2.21

Line 322: How is it determined that the fit works well?

This is done visually, as mentioned in line 309. The quality of each filter has to be verified visually.

R2.22

Basically, the power law method has a gradient threshold that depends on the concentration.

This is correct and described in section 2.4.1.

R2.23

Line 338: NPF events can exceed this threshold. If the authors mean that during the deployment no NPF events exceeded this threshold, then evidence should be provided for this assertion.

This is true, they could exceed this threshold. Therefore it is important to be able to adapt this threshold to different datasets. But in our dataset from MOSAiC they never did. The NPF event on June 21st in Figure 5c was the event with the highest concentration.

R2.24

Line 349: This should be stated above when the PDA is described, not after an example. However, this sentence is not clear, as there are separate threshold and gradient filters.

We agree, this needs to be addressed more precisely. We answered this in R1.3. In summary, the threshold filter is applied simultaneously with the derivative filter. We therefore implemented the section on the threshold filter as a new paragraph to the section 2.4.1 about the derivative filter.

R2.25

Line 358: The statement (on line 361) that application of this filter discards points is true, but realistically how many points (of 10-s duration) will be lost by doing this?

As shown in Table 2, the neighboring points filter flags another four per cent of the data points after the application of the derivative filter. The sparse filter only flags another additional 1%. To answer R1.11 we extended Table 2 in section 3.1 where we compare how strong a change in each filter's parameters affects the dataset.

R2.26

Line 367: This implies that pollution can occur in individual 10-s intervals. How are individual spikes in 10-s data points determined to be pollution and not issues such as someone bumping the instrument or an electronic glitch, which sometimes occur?

Here we refer to false negatives (polluted data points which are not flagged). For instance, if three data points in a row show a small derivative, they will not be detected by the derivative filter. This happens not often, but it happens. Therefore we implemented the median filter, which detects those leftovers. Someone bumping the instrument, or an electronic glitch, could also lead to outliers, which would potentially also be detected by this step, as they cannot be distinguished from contaminated data points. In both cases, we would not want these data points to contribute to our background signal. We changed the first sentence in Section 2.4.3 Step3: Median filter to mention that we address false negatives here:

The median filter aims at detecting false negatives, i.e. data points which are not representative of the background signal but were not flagged by the previous filter. For each data point, we calculate its deviation from the running median over a time interval (the median time interval).

R2.27

Line 391: This shows that the PDA algorithm works, but for this example it would seem that a threshold or median filter would work equally well.

We address this in the answer to R2.4. In short, a single threshold or median filter would not perform as well as the combination of all filters in the PDA and most importantly, a derivative filter is needed as a first step. As mentioned before, the added value of the PDA is that a single set of parameters can be applied to clean an entire time series (rather than selecting case-dependent parameters).

R2.28

Line 397: The color scheme in Figure 4d makes it difficult to determine which region is which. Also, this figure should be for only the time period of Figs. 4a, 4b, and 4c. It appears that Figure A6, for a different day, has the same panel d as Figure 4. Figure A6 is another example where the algorithm doesn't do much better than merely filtering by eye – it is too easy of an example to illustrate the utility of the PDA.

Thank you for pointing this out, we changed the colors of the histograms in Fig. 4d and A6d. However, the two examples show only 24h of data, and therefore a histogram of such a short period would not be representative of the performance of the PDA. To illustrate the functionality of the PDA it is necessary to show case studies of a short period of 24h, so that one can identify single data points to see how individual filtering steps work. As mentioned previously, we do not aim to provide a better detection performance than visual detection, but the point is that filtering > 2.8 Million data points by eye is highly time-consuming.

R2.29

Line 403: This statement seems odd, as further filtering would not have removed any other points, so the claim that this "allows retaining more data" does not seem justified, and seems to contradict the previous statement.

Thank you for this comment. What we meant is that in some cases we found that the derivative filter did already detect all contaminated data points, so no more additional filters are needed. We agree, this sentence might be confusing, so we decided to remove it.

R2.30

Line 406: This statement doesn't justify use of the PDA, as simply filtering by wind would retain roughly the same number of points. A real comparison would be to show which points the PDA removes that the wind filter doesn't, or vice-versa. All that is being demonstrated here is that if the gradient filter is used first, the other filters have little additional effect.

This line says that the derivative filter already flags the majority (56%) of data points, which demonstrates that this filtering step is very efficient, and the following steps only contribute a small amount to the flagged data points. Adding the wind filter here might be useless, we deleted this column. A comparison with the wind filter is demonstrated in Section 3.2.1 and Fig. 8. There, we see that the wind filter leaves many false negatives, compared to the PDA.

R2.31

Line 410: Figure A7 should be presented as a time series so the reader doesn't have to scan up to down sequentially. The values tell nothing about the PDA, only about the data during this cruise.

The purpose of Fig A7 is to sum up the whole expedition data after application of the PDA. It is meant as an overview example. We used the same format as in Angot et. al. (in prep). This figure is indeed not essential here (reason why it is included in the appendix) but of high interest for the MOSAiC community.

R2.32

Line 435: In Figure 5a it appears that 6 points, corresponding to roughly one minute of data out of the entire day, was removed, and these would have easily been removed by a median filter. Similarly for 5c, and in 5b a median filter would detect the onset. A smoothing algorithm would have removed the spikes.

Yes, a median filter might have detected some of these data points, but not all of them. We still need the full PDA to detect the contamination as demonstrated in the answer to R2.4. Individual filtering steps would work in specific situations, but only the combination of all of them allows to detect contamination in a big dataset like the one we have from MOSAiC.

R2.33

Line 452: This is not an especially compelling result, as the decrease does not appear extremely abrupt.

This is true, and because the decrease does not appear very abrupt, the PDA works. Still if we would not show this example, the reader could ask: "What happens during washout events, or during NPF events, where we have a decrease or an increase in particle number concentration within a relatively short amount of time?". For this reason we think that providing the two examples in Fig. 5c and 5d is adequate.

R2.34

Line 457: The authors should verify that the spikes ARE caused by pollution, not that they are ASSUMED to be.

Line 457: Again, this is not particularly compelling, as the NO shows only a minor increase at 12h on July 27, whereas the number concentration shows a very large increase, but the CO shows none. This would seem to require more explanation of possible sources that would increase NO and number concentration but not CO.

CO does not react strongly to ship pollution as the one year long dataset of MOSAiC (and other ship expeditions) demonstrates. It is hence not a good tracer for ship exhaust. On the other hand, the BC data and the NO data support the CPC data and the assumption that this spike is a pollution spike. The wind direction during this period is close to 180°, which is the direction from the stack and is another indicator that these data are influenced by local contamination. We added the sentence:

Note, that the CO signal does not react strongly to ship pollution.

R2.35

Line 484: The statements that this "validates the functionality of the PDA" and that it "shows the ability of the PDA to detect pollution in datasets with different time resolutions" seems a bit overblown. Additionally, the time period between 02/19 00h and 02/21 00h shows a very large increase (nearly an order of magnitude) in particle number concentration, yet the argument is that this is background because it is not filtered out by the PDA. The average size is also quite large (as shown by the yellow shaded region). Some discussion of the source or composition of this aerosol seems to be required to demonstrate (or at least argue) that it is not a well-mixed aged polluted plume.

We have to clarify that the purpose of the PDA is not to detect well-mixed aged polluted plumes. On the contrary, we aim to retain these plumes, because they constitute an important feature of the Arctic atmospheric composition. We want to detect fresh, locally created pollution from the ship or surrounding sources (see the answer to R2.9, where we modified the text in section 2).

To demonstrate that the large increase on the 19th of February is actually caused by a warm air mass intrusion carrying contamination from northern Russia, we show a figure below. It shows on the left panel the temperature (in blue) and the particle number concentration (in black) and the SO_2 mixing ratio (in red), during this day. We see that the particle number concentration rises together with the SO_2 mixing ratio and that the whole period was accompanied by warm temperatures intruding into the Arctic from further south. On the right panel we see the air mass transport as simulated by the Lagrangian particle dispersion model FLEXPART for the 19th of February (https://srvx1.img.univie.ac.at/webdata/mosaic/mosaic.html). The air masses at Polarstern were coming from the region of Norilsk, RU, a known source region for SO_2 (Sipilä et al., 2021).

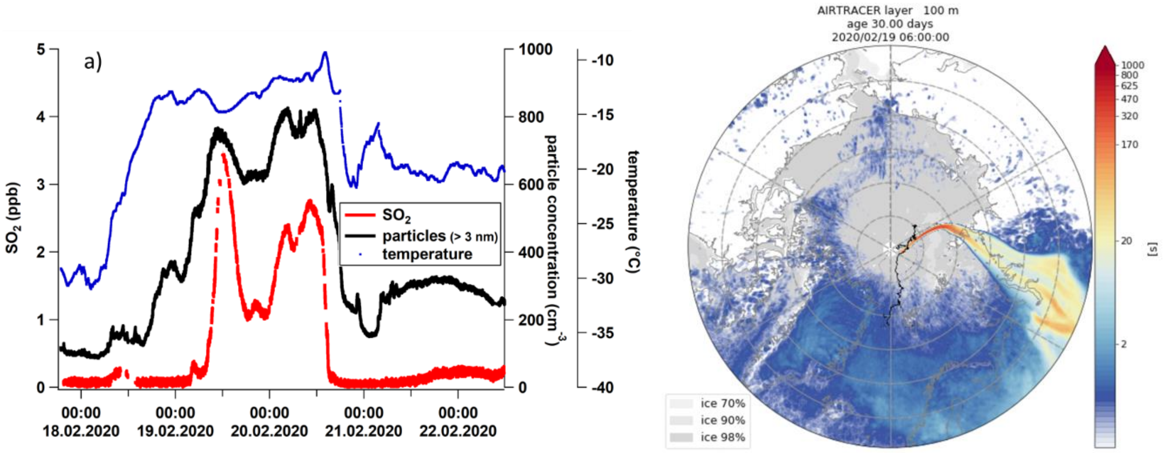

Figure R2.35.1. SO_2 mixing ratio (red), particle number concentration (black) and temperature (blue) during a warm air mass intrusion event on 19 February (left panel) and the back trajectories from FLEXPART for the same day (right panel).

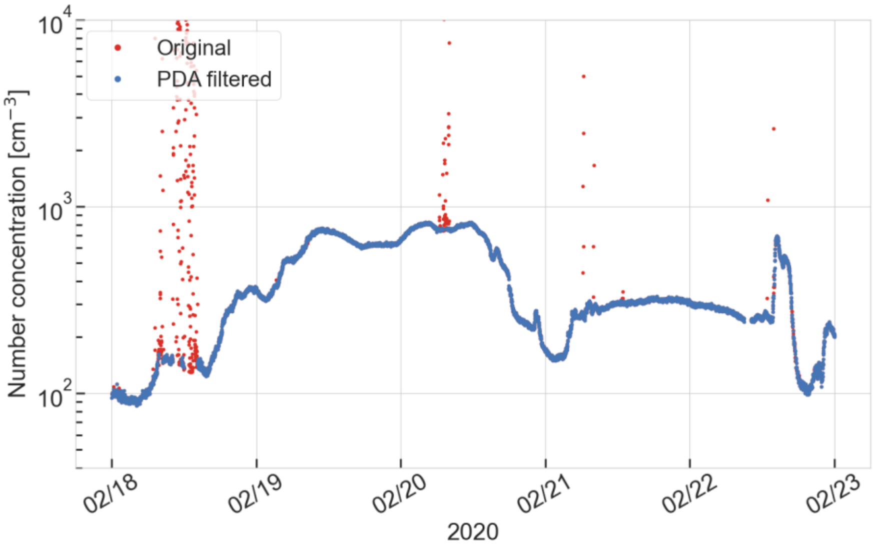

Figure R35.2. Particle number concentration from February 18-22, measured by the CPC3025. Data identified as polluted are marked in red, data filtered by the PDA in blue.

The second figure shows the PDA applied to the CPC on the case study of February 19th. Polluted data points are marked in red, the retrieved signal is shown in blue. The PDA works in retaining high particle number concentrations that are not caused by primary pollution.

R2.36

Line 496: The fraction of data marked as polluted by the PDA should be given for comparison.

We added it.

R2.37

Line 496: It seems as though a median filter would remove the points shown in panel a of Fig. 8.

We chose this case study in Figure 8 because during exactly this time, a member of our team was present in the measurement container and saw how the CPC reacted on the snow groomer passing by several times. We compared the different filtering steps in the answer to R2.4 where we show how efficient single filtering steps are. Even if the median filter alone works in some single case studies, it does not work for the entire dataset. For this reason we use the PDA.

R2.38

Line 500: The statement that “the PDA detects all polluted data” requires justification. That is the hypothesis that the manuscript is trying to justify.

We changed the statement “the PDA detects all polluted data” to “the PDA detects more polluted data than the wind filter”.

R2.39

Line 504: comma should be removed

We removed it.

R2.40

Line 514: The choice of terminology "case-sensitive" is odd and should be replaced by something more descriptive.

We removed the term "is not case-sensitive and thus" in the caption.

R2.41

Line 522: Displaying figure 9 on a logarithmic scale makes it difficult to visualize the magnitudes. A more illustrative method would be to show a section of a time series that compares the PDA results with those from the visual approach.

We show exactly this in Figure A8. Still, Figure 9 allows to visualize the total dataset, whereas in Figure A8 we only show a part of it. We changed the time range (5 days instead of 31) and the colors in Figure A8 for better readability.

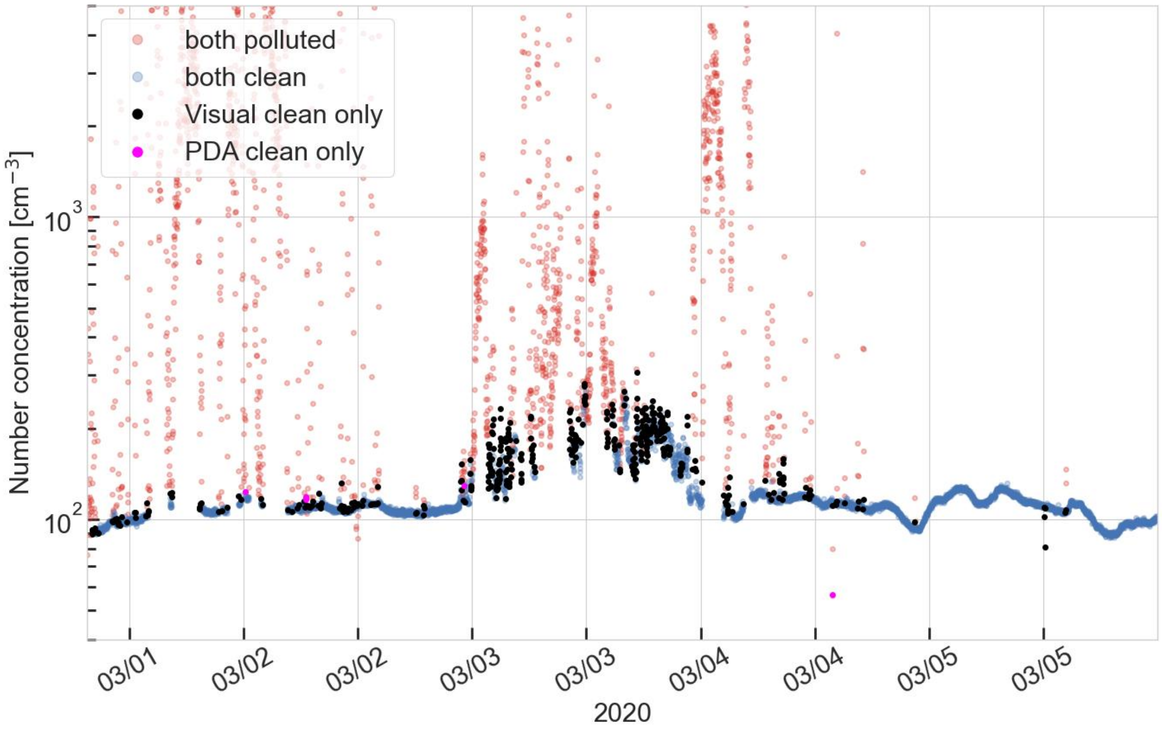

Figure A8: Time series with a comparison of the visual identification method and the PDA between 1 and 5 of March. In red: Data points which are detected as contaminated by both methods. In blue: Data points which are detected as unaffected from pollution by both methods. In black: Data points which are detected as unaffected from pollution only by the visual identification method. In magenta: Data points which are detected as pollution-free only by the PDA.

R2.42

Line 528: Basically, the arguments are that the PDA is easier to apply and that it is more objective, but it is not clear that these arguments are valid. Yes, it would be a lot of effort to apply the visual method to a year's worth of data, but it would be done once and could be applied to all relevant data sets. The subjectivity argument needs to be justified by comparing different thresholds and values in Table 1, which are subjective by nature.

We would like to note that one could clean the CPC MOSAIC dataset by eye once, but that would not help to clean datasets from other campaigns or other instruments. The point of the PDA is to have a tool for many campaigns, and moreover long-term continuous monitoring stations such as the Jungfraujoch, where repeated visual cleaning is time consuming. We see a great advantage in the PDA.

We answered the question regarding objectivity in R1.1 and dedicated a paragraph in the introduction to this. We add a copy of the answer to R1.1 here:

The reviewer is correct that setting between 3 and 7 parameters is not entirely objective. However, the treatment of the dataset in itself is consistent and reproducible, because these parameters are not changed. All parameters are set once in the beginning and applied to the whole dataset. Hence the algorithm is more time efficient and more consistent in itself in terms of strictness, compared to, for instance, a manual cleaning method, where the expert has to decide on each individual data point. The 3 to 7 parameters also make the algorithm flexible. We aim to provide an algorithm, which can be used with many different remote atmospheric datasets. This comes with the trade-off that the user needs to adjust all parameters. In other words: If we restrict the parameters, we exclude more datasets and more use cases, hence the algorithm could not be applied to many other user's datasets. To emphasize the advantages of the PDA, we have made the following change in the manuscript in section 1, as mentioned in R1.1:

This makes the algorithm an efficient and consistent way to detect local contamination in large remote atmospheric time series, as they exist for example from ship campaigns or from remote stations. This method is objective as the treatment of the data is consistent throughout the whole time series considered, because the same value of each parameter is applied to the entire dataset.

R2.43

Line 530: The color scheme and small size are such that it is difficult to evaluate the comparison. It would be better to show a small fraction of the time series, such as 03/02 to 03/04 where the two methods might differ.

Thank you, indeed the colors were not easy to read. We changed the colors of Fig. A8 and shortened the time period to 5 days (March 1st to March 5th). This is shown in R2.41.

R2.44

Line 555: Separation on the basis of the gradient might not be possible, but the two branches are clearly quite separable by a threshold. Although this strong threshold doesn't agree exactly with the results by the PDA, the authors should list the number of cases that they agree and that they don't so a valid comparison can be made. The statement that this threshold "failed to produce a reliable pollution mask" is not justified; I would argue that if it got 95 % of the cases and missed 5 %, it would be rather reliable. The authors are taking their PDA as the gold standard, but this assertion has not been justified.

Line 558: The fact that the authors need to hypothesize about the failure of the gradient method seems to undercut their earlier assertions that the PDA can be applied to any time series.

Line 562: Again, if another approach is required to produce a pollution tag for AMS data, then this demonstrates restrictions on the general applicability of the PDA.

We reply to the above three questions together.

Pollution detected by the AMS is of different nature than what the CPC sees. It reacts less sensitive to fresh pollution because it has a lower cut-off size of 70 nm, which is larger than the peak of fresh contamination at 30 nm. We also found that to clean the AMS data from pollution, one simple tracer (in our manuscript we chose m/z = 57) did not work to detect all fresh contamination influence. This is because the same markers (not only the one chosen here) also occur in more aged and long-range transported air masses. Therefore, we could not apply the PDA to the AMS and another detection method, based on the mass spectrum of the measurement, was developed for the AMS data (Dada et al., in review).

We tested the PDA successfully on particle number concentration data and on trace gas data and point out that it is meant to be applied to primary pollution data. This section is meant to demonstrate that the PDA does not work with any time series. For example, it does not work on accumulation mode particle chemical composition data, which contains secondary particles, such as the AMS data.

R2.45

Line 572: This is another instance in which the gradient filter does not work well. A simple threshold of ~420 ppm would seem to work quite well, as for physical reasons there would be few situations where higher values of CO2 mixing ratio would not be from pollution.

A simple threshold in the CO_2 data would leave too many false negatives or flag too many false positives, as shown in the attached figure below. We added a simple threshold line at 400 ppm (blue horizontal line). This threshold would not capture all contaminated data points. Also, the CO_2 concentration varies strongly with the seasonal cycle, as stated in R2.3:

The same applies for the CO_2 time-series; the strong seasonal cycle (419.9 $\pm$ 1.5 ppm in January (mean $\pm$standard deviation) vs. 401.7 $\pm$ 1.1 ppm in August) makes it impossible to apply a simple threshold for the whole dataset.

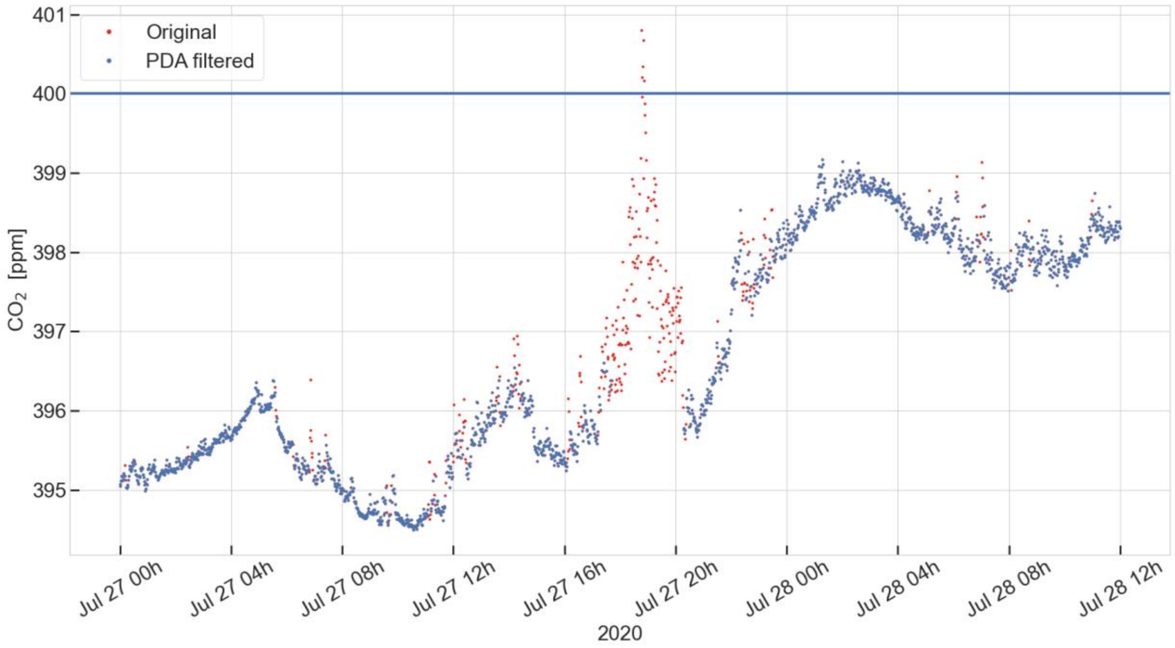

Figure R2.45. CO_2 concentration during July 27–28. Data identified as polluted are marked in red, data filtered by the PDA in blue. The blue line represents a cut-off threshold line at 400 ppm.

R2.46

Line 573: The reader at this point does not remember what “step 1B” is, so please state it explicitly. If it is merely the threshold value, then this is a threshold filter.

Step 1B is the derivative filter step based on the IQR method. We added a reminder of step 1B to the reader:

We thus applied the PDA with step 1B (the derivative filter based on the deviation from the running interquartile range) to the CO_2 dataset.

R2.47

Line 580: The discussions of the m/z=57 and the CO2 time series demonstrate that the gradient method did not work, and that a simple threshold method did.

As shown in R2.45, a simple threshold would not work to detect all polluted data points in the CO2 time series. Neither would it work for the CPC time series. As demonstrated in R2.4, all steps of the PDA are needed to detect the data points which are affected by local contamination.

R2.48

Line 590: This is yet another instance in which it seems that a median filter would remove the same points. Additionally, there is a time near 07-18-16 h when values in blue are much below the others, and any visual inspection would remove these as anomalous, but they weren’t removed by the PDA.

A median filter would not work on this dataset, but the derivative filter based on the interquartile range does. The mentioned values of low concentration could indeed be missed outliers by the PDA. The PDA is not perfect, of course, but it detects most of the polluted data points and compares well to the visual cleaning method.

R2.49

Line 602: Again, it looks as though a median filter would work well and give essentially the same results.

Please refer to R2.4 and the dedicated discussion. To demonstrate how a single threshold and a single median filter would perform on this example, we added the figure below. It shows the original time series in red and the blue dots show the retrieved data after the application of a) only a derivative filter with a = 0.5, m=0.55, b) only a threshold filter, where we chose an upper threshold of 10000 and c) only a median filter, with a median window of 1h and a median threshold of 1.1.

The derivative and the median filter perform quite good in this example. But as demonstrated in R2.4, a median filter does not work alone in many other situations. In contrast, the derivative filter works well throughout the entire dataset, and in combination with other filters, the PDA delivers satisfactorily results.

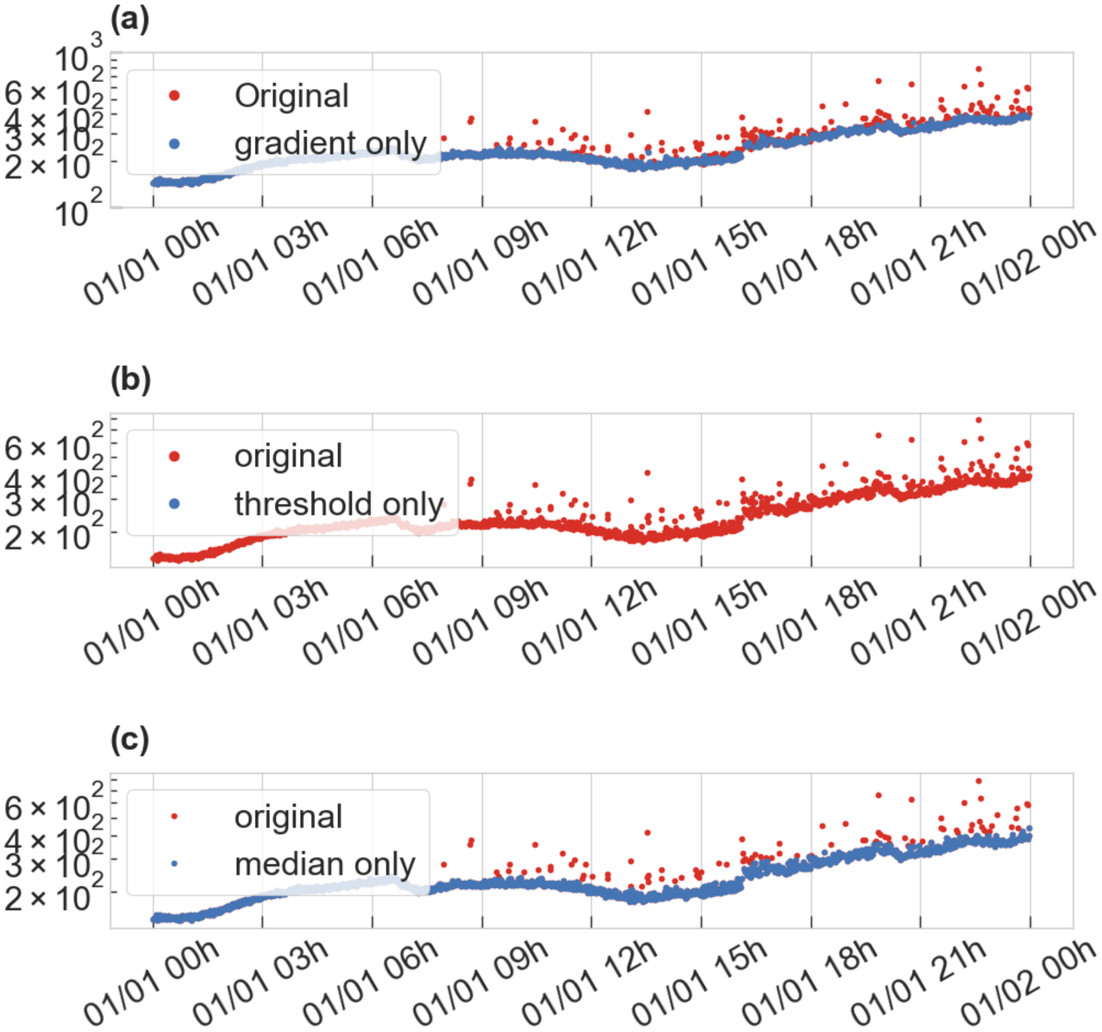

Figure R2.49. Case study of particle number concentrations on 1 January, filtered by a) the derivative filter, b) the threshold filter, c) the median filter. Data identified as polluted are marked in red, data filtered by the gradient filter in blue.

R2.50

Line 590: The term "baseline" is not an appropriate one here. Presumably the authors mean "background."

We agree, and changed the word baseline to background.

R2.51

Line 604: Low pollution would be very difficult to detect, with any time series method. This is why other data streams, such as CO and BC are often used.

We agree that ancillary data is the best way to verify whether a certain time window is polluted or not. But it is not always available. Please, see our dedicated discussion on this in R2.8.

References:

Angot, H., Blomquist, B., Howard, D., Archer, S. D., Bariteau, L., Beck, I., Boyer, M., Brasseur, Z., Helmig, D., Hueber, J., Jacobi, H.-W., Jokinen, T., Laurila, T., Posman, K., Quéléver, L. L. J., Shupe, M. D., and Schmale, J.: Year-round trace gas measurements in the Central Arctic during the MOSAiC expedition, submission to Scientific Data, MOSAiC special issue, in prep.

Brantley, H. L., Hagler, G. S. W., Kimbrough, E. S., Williams, R. W., Mukerjee, S., and Neas, L. M.: Mobile air monitoring data-processing strategies and effects on spatial air pollution trends, Atmos. Meas. Tech., 7, 2169–2183, https://doi.org/10.5194/amt-7-2169-2014, 2014.

Dada, L., Beck, I., Quéléver, L. L. J., Baccarini, A., Angot, H., Laurila, T., Brasseur, Z., Boyer, M., Jozef, G., De Boer, G., Henning, S., Daellenbach, K. R., Jokinen, T., and Schmale, J.: A central Arctic extreme aerosol even triggered by a warm air mass intrusion, Nat. Comm., under review

Gallo, F., Uin, J., Springston, S., Wang, J., Zheng, G., Kuang, C., Wood, R., Azevedo, E. B., McComiskey, A., Mei, F., Theisen, A., Kyrouac, J., and Aiken, A. C.: Identifying a regional aerosol baseline in the eastern North Atlantic using collocated measurements and a mathematical algorithm to mask high-submicron-number-concentration aerosol events, Atmos. Chem. Phys., 20, 7553–7573, https://doi.org/10.5194/acp-20-7553-2020, 2020.

Giostra, U., Furlani, F., Arduini, J., Cava, D., Manning, A. J., O'Doherty, S. J., Reimann, S., and Maione, M.: The determination of a "regional" atmospheric background mixing ratio for anthropogenic greenhouse gases: A comparison of two independent methods, Atmos. Environ., 45, 7396–7405, https://doi.org/10.1016/j.atmosenv.2011.06.076, 2011.

Hagler, G. S. W., Lin, M.-Y., Khlystov, A., Baldauf, R. W., Isakov, V., Faircloth, J., and Jackson, L. E.: Field investigation of roadside vegetative and structural barrier impact on near-road ultrafine particle concentrations under a variety of wind conditions, Science of The Total Environment, 419, 7–15, https://doi.org/10.1016/j.scitotenv.2011.12.002, 2012.

Liu, J., Dedrick, J., Russell, L. M., Senum, G. I., Uin, J., Kuang, C., Springston, S. R., Leaitch, W. R., Aiken, A. C., and Lubin, D.: High summertime aerosol organic functional group concentrations from marine and seabird sources at Ross Island, Antarctica, during AWARE, Atmos. Chem. Phys., 18, 8571–8587, https://doi.org/10.5194/acp-18-8571-2018, 2018.

McNabola, A., McCreddin, A., Gill, L. W., and Broderick, B. M.: Analysis of the relationship between urban background air pollution concentrations and the personal exposure of office workers in Dublin, Ireland, using baseline separation techniques, Atmospheric Pollution Research, 2, 80–88, https://doi.org/10.5094/APR.2011.010, 2011.

Sipilä, M., Sarnela, N., Neitola, K., Laitinen, T., Kemppainen, D., Beck, L., Duplissy, E.-M., Kuittinen, S., Lehmusjärvi, T., Lampilahti, J., Kerminen, V.-M., Lehtipalo, K., Aalto, P. P., Keronen, P., Siivola, E., Rantala, P. A., Worsnop, D. R., Kulmala, M., Jokinen, T., and Petäjä, T.: Wintertime subarctic new particle formation from Kola Peninsula sulfur emissions, Atmos. Chem. Phys., 21, 17559–17576, https://doi.org/10.5194/acp-21-17559-2021, 2021.