Comment on amt-2021-429

Anonymous Referee #2

Referee comment on "Automated identification of local contamination in remote atmospheric composition time series" by Ivo Beck et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-429-RC2, 2022

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.

GENERAL COMMENTS

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). 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.

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 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.

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.

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.

MINOR COMMENTS

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.

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.

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

Line 263: stated earlier on line 105

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.

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

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

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

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?

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

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.

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

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.

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

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

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.

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.

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?

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?

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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

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

Line 504: comma should be removed

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

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.

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.

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.

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.

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.

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.

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.

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.

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

mean "background."

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

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.