Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-450-RC2, 2020

Atmospheric Measurement Techniques
Discussions

Interactive comment on "Filtering of pulsed lidars data using spatial information and a clustering algorithm" *by* Leonardo Alcayaga

Anonymous Referee #2

Received and published: 12 April 2020

Alcayaga presents a study about filtering methods for Doppler wind lidar measurements. A new method based on data clustering is developed and compared against the classical CNR filter and a median filter which has become more popular recently. The method is tested in a simulation with artificial turbulence and noise as well as in a real experiment. I think the method is promising and the results that are shown look very interesting. However the manuscript is way too long, not prepared very well and should be rewritten in a much more concise way. The structure currently is confusing with many repetitions and lengthy explanations of minor details, but important information about the data, the methods and the results are missing. Since the topic of the study is relevant and the methods and results could be interesting for the scientific community I would like to see a major revision of the manuscript before it could be

reconsidered for publication in Wind Energy Science. I give general comments about each section as well as specific comments in the following.

0.1 General comments

- It has not been shown convincingly that the generated noise in the lidar simulation is realistic and the analysis of the filter in the simulation can thus be considered relevant for real-world measurements.
- The math of the methods is not presented clearly in equations, especially regarding the filters.
- The work is not referencing important work in the field of lidar simulation and daa filtering adequately.
- Section 3.2: Lidar simulators are not new and similar work can be referenced (e.g. Stawiarski et al. 2013, Gasch et al. 2020). Based on these works, the description of the technology could be siginificantly shortened. The most important points like the resolution of the synthetic data that is used should be highlighted in a concise way.
- Section 3.3: The noise generation is described with many words and steps that are very hard to follow and confusing. I think it should be possible to describe a noise filter transfer function with a concise mathematical expression. I also think that in this section the characteristics of the synthetic noise should be compared to what is expected from real lidar measurements. Could you for example show a PDF from real measurements of only low CNR data in comparison to the artificial noise? Without any information on how realistic the synthetic noise is, it is hard to judge the quality of the filter from the simulation results.

- Section 4.1 is partly a repetition of things that have been said in the introduction and since CNR-filters are very easy and well known, I think this could be cut much shorter.
- Section 4.2 is supposed to describe the median filter, but does not give the most important parameters. The median of what database is used? Just single scans, multiple scans, the whole scan or just parts of it. Again, I recommend to put the filter description into one or two equations, which would describe it in the best concise way. Menke et al. 2019 and Menke 2020 (dissertation) introduced a modified three-stage median filter for spatial scans. How does the method applied here relate to that?
- Section 4.3 gives a lengthy description of the clustering algorithm, but misses the most important point. Where is the connection between the lidar parameters CNR, Vlos etc and the filtering algorithm. Please give the filter functions for the concrete problem of lidar signals. What is the k-distance function fo the lidar measurement? How is the data sorted in Figurer 8? I doubt that any lidar user can reproduce this method with the information which is given in this section.
- Section 5.1: The author introduces many performance metrics here, of which many are not very useful in my opinion and only add to the confusion of the reader. To me, the interesting metrics are the fraction of good observations (here: η_{recov}) and the false positive rate (i.e. the percentage of data points that are considered good observations although they are contaminated by noise).
- Section 5.2: I would advice the author to focus on just one most appropriate metric for the analysis of the similarity of the PDFs, especially since the qualitative results are the same and differences between the two metrics are not discussed in Section 6 and 7.
- Section 6.1: I think the the line-of-sight threshold should be discussed in Section

4.2 and not here. What I miss in this section is a plot of the actual LOS velocity fields recovered with the two filters. Lines 403ff give a discussion that is partly repeated in Section 7.1 and should be removed here.

- In section 6.2 the author argues a lot with data recovery, which is not a good metric, because without any filter, the data recovery is perfect, but includes a lot of bad data. The author should focus on the metrics introduced in section 5.2, which is a good choice and the best that can be done. So, I wonder if Figures 15-17 and Table 5 are really useful for the study. One idea would be to replace Figure 16 with a plot of the PDF of the area around the hard target only, comparing the three filters and the original data. Same as for data in different distances to the lidar.
- I think the title "performance assessment" of sections 7.1 and 7.2 is misleading, because those sections mostly evaluate the flaws of the test cases. The performance of the filters is already assessed in the results section.
- Section 7.3 and 8 could probably be combined.

0.2 Specific comments

- p.1, l.1: simultaneous multi-point observations are possible with masts if multiple sonics are installed.
- p.1, l.2: write "lower" instead of "reduced"
- p.1, l.4: "reduced data recovery" compared to what? I am also not sure if "data recovery is the proper term.
- p.1., l.6: "...spatial position, and V_{LOS} smoothness". The abstract needs to be understood without reading the whole manuscript. It is not clear at this point what is meant by spatial position and smoothness.

- p.1,l.13: "its adoption" - "their acceptance"!?
- p.1.,l.21: Since the CNR thresholds are so divers and depend on the conditions and instruments I recommend to not give numbers here.
- p.2.l.37: typo "approaches"
- p.2,l.56: "DBSCAN" acronym should be explained here.
- p.3,l.80: Why are the scanning patterns coherent?
- p.5,l.105: The term "numerical lidar" is very unusual and irritating. I would recommend "lidar simulator" or "virtual lidar".
- p.5,l.112: What does "coarse" mean here? Numbers should be given.
- p.6,Eq.2: The variable names are somewhat confusing, because what is here Δp is ΔR in the references of Smalikho and Banakh and Δp in the references is r_p here.
- p.6,l.129: "corresponding range gate center"!
- p.6,l.130: "range gate length" is not very specific. If you give the explanation of r_p from FWHM, you could also give the explanation of Δp from the time window of the FFT.
- p.7,l.149: typo "radial"
- p.7,l.154: referencing Figure 10 which is introduced much later, is bad style.
- p.7,l.158: type "in"
- p.8,l.177: again, a figure (Figure 5) is referenced before its introduction.

- p.9,l.180: "The fraction of beams contaminated at each band..."
- p.9,l.183: typo "from".
- p.10,l.201: I do not think you can really give a common value for CNR values. They depend strongly on instruments and location.
- p.10,l.215f: put citation Huang et al in parantheses.
- p.13,l.251: the m in "m-dimensional" is not explained.
- p.14,l.285: How does $d_k(n)$ look like for the lidar signal problem?
- p.15,l.298: typo: "noisy"
- p.15,l.298: Figure 9 is referenced before introduction.
- p.15,l.302: Equation 6 is referenced before it appears. Please introduce it before.
- Figure 8b) seems to be moreless the same as Figure 5b.
- p.17,l.333: "PDF" should be in capital letters as an abbreviation.
- p.18,l.344: Something is wrong with the grammer in this sentence.
- p.18,l.345: What is the value of α that is used in this study?.
- p.18,l.345: Again, grammar.
- p.18,l.345f: The numbers about the amount of data that was analyzed should be given in Section 2.
- p.19,l.372: remove one "then".
- p.20,l.386: typo: "account"

- Figure 11: I think this figure is not neccessary. If it is still shown, labels have to be larger.
- Figure 14: typo, should be "phase 2"
- Figure 15: Why is no upper threshold for the CNR filter applied, which would remove the wind turbine hard target from the recovered data?
- Figure 16: I think this plot is not neccessary.
- p.23,l.439: What is meant by "quality of the data"? Probably you mean a lower false positive rate, but how do you know?
- p.23,l.443: Metrics are introduced in Sect. 5.1.
- p.23,l.443f: Again, quality is undefined.
- p.27,l.502: typo: "from"

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-450, 2020.