Comments on ”Filtering of pulsed lidars data using spatial information and a clustering algorithm”

Anonymous Referee #1

General comments

This paper presents alternative wind lidar data recovery methods, over the traditional carrier-to-noise ratio. The paper presents both a clustering technique and a median-like filter, and evaluates results on both synthetic and real lidar data.

While the paper includes some important results, the presentation is a little clumsy, and I feel the paper could be greatly improved. There needs to be general improvements to the usage of English throughout, examples of which I have highlighted below. The paper overall reads as if several authors composed different sections, there is a lot of repetition of the discussion, and the figures do not flow nicely. While some scrolling/page turning is expected, referring to figure 10 on page 7 requires the reader to turn to page 19. Perhaps there is an alternative way to make your point on page 7? Figure 7 also does not seem to be referred to in the text?

A: The paper will be improved to correct the problems suggested by the reviewer.

The point I would like to make most clearly is your conclusion states the clustering filter performs best in both synthetic and real data, and increases data availability between 22% and 38%, while also reducing erroneous measurements between 70% and 80%. This is a significant result, and I feel you could make more of this in the paper. There is a lot of discussion on methods used, sometimes repeated several times, but I feel comparatively little on your major results. Improving the flow of the paper, and removing some of the repeated discussion to focus more on results will greatly enhance your paper.

A: This is an important comment and will be reflected in the corrected version.

Specific comments

1) Title should read “lidar” rather than “lidars”
A: The work presents results from a real and a lidar simulator, this is the reason behind lidars instead of lidar.
2) Page 1 Line 13 – replace “its adoption” with “their adoption” or similar Line 14/15/16 – the meaning of the sentence beginning “Their capability to measure: : :.” is unclear. Do you mean a single lidar can scan a spatial domain of comparable size to a wind farm? If so, it would be helpful to include an indication on the actual size of a windfarm By “their increasing accuracy” do you mean increased accuracy over meteorological masts? Line 17 – please be more specific with “traditional wind measurement techniques”, for example wind profiling radars can also be used, and are also susceptible to atmospheric conditions. What is “traditional”? Line 18 – please define “lack of references”, do you mean a second instrument to compare wind values to? Line 25 – please define VLOS the first time you use it, rather than the second
A:
- Page 1 Line 13. Corrected.
- Page 1 Line 14/15/16. Since wind farm vary in installed capacity and size, giving one number it is not easy. But spacing of large turbines can be in the order of a kilometer (assuming 6D streamwise spacing for turbines of around 150m rotor diameter, the long-range scanners here can cover up to 7km x 10km, meaning several turbines), for clarity, the

sentence was deleted. The technology has developed the last years to increase the laser energy and the backscatter signal quality.

- Page 1 Line 17. The sentence was corrected to refer to meteorological masts.
- Page 1 Line 18. Yes, clarified in the text.
- Page 1 Line 25. Corrected.

3) Page 2 Line 26 – remove the "of" in "between of line-of-sight: : :." Line 39 – you don't need both "like" and "e.g." together Line 39 – please consider rephrasing the sentence beginning "Complementing all these features: : :.". The sentence is very long and difficult to follow. Line 45 - ": : :.which are capable of classify large data sets: : :." needs to be reworded for correct English Line 54 – swap the order of "defines" and "always" to read "which always defines a unique: : :." Line 56 – please define/introduce DBSCAN here, rather than on page 12 Line 58 - ": : :.capable of identify clusters: : :." should read ": : :.capable of identifying clusters: : :."

A:

- Page 2 Line 26. Corrected.
- Page 2 Line 39. Corrected.
- Page 2 Line 45. Corrected.
- Page 2 Line 54. Corrected.
- Page 2 Line 56. Corrected.
- Page 2 Line 58. To keep the introduction section short, DBSCAN definition needs its own section, this is mentioned in the corrected version.
- Page 2 Line 58. Corrected.

4) Page 3 Line 72 – what do you mean by "the wind speed data covers a large horizontal area"? Do you mean you wish to measure winds across a large area? Line 88 – I'm not sure I follow what a "wrong observation" is, as compared to an outlier?

A:

- Page 3 Line 72. It will be rephrased for clarity.

- Page 3 Line 88. It will be rephrased for clarity.

5) Page 5 Line 99 – change "generate" to "generates" Line 102 – change "make" to "mean" or similar.

A:

- Page 5 Line 99. It will be corrected.

- Page 5 Line 102. It will be corrected.

6) Page 7 Figure 2 caption – line 3, I believe should read "next" not "nest" Line 149 - "radial" is miss-spelled Line 158 - "en" should be "in".

A:

- Page 7 Figure 2 caption. It will be corrected.

- Page 7 Line 149. It will be corrected.

- Page 7 Line 158. It will be corrected.

7) Page 9 Line 184 - “2” should read “section 2” as done previously Line 189 – the sentence beginning “The noisy areas show: : :.” is very long and hard to follow. Please consider rewording.
A:

- Page 9 Line 184. It will be corrected.

- Page 9 Line 189. It will be rephrased for clarity.

8) Page 10 Line 200 to 203 – these 2 sentences seem to be a repeat of the introduction?
A:

- Page 10 Line 200 to 203. Sentences It will be eliminated.

9) Page 11 Line 229 - “non” should read “not”
A:

- Page 11 Line 229. It will be corrected.

10) Page 12 Line 240 – similar to the comment above, page 10 lines 200 – 203, this section appears to be a repeat of earlier discussions
A:

- Page 12 Line 240. It will be rephrased.

11) Page 15 Line 298 – I think you mean “noisy” not “nosy”
A:

- Page 15 Line 298. It will be corrected.

12) Referral to figure 7?
A:

- Referral was only in the caption of Figure 8, which is complementary. It will be corrected in the text.

13) Page 16 Lines 315 to 320 – sentence beginning “This allows us to define: : :.” is very long and difficult to follow Line 320 - “this metrics” should read “these metrics”
A:
-Page 16 Line 315 to 320. It will be rephrased.

14) Page 18 Line 344 – I think you are missing “are” in “: : :.that two realizations from the same distribution: : :.” Line 365 – should read “: : :.on the other hand: : :.” rather than “in”
A:
-Page 18 Line 344 and 365. It will be corrected.

15) Is there a reason why you can’t do the same tests to the synthetic data as you are for the real data?
A: It is possible, but tests on real data are based on reliable observations on a range of CNR values, due to the lack of references available, which is not the case for synthetic data. In the ideal case, the test applied on synthetic data would be the best for real data.

16) Page 19 Line 372 – remove the second “then” from “: : :.then becomes relevant then: : :.”
A:

-Page 19 Line 372. It will be corrected.

17) Page 20 Line 387 – remove the comma after “both” to read “: : :.in both noisy and reliable: : :.” Line 390 – reverse the order of “be then” to read “then be” Line 391 – replace “its” with “their” to read “: : :.distant from their previous location: : :.” Line 401 – remove “be” and change “benefited” to “benefit” to read “: : :.filter will benefit by: : :.” Line 403 – add “to” to read “: : :.dimensions to the data description.”
A:
- Page 20 Lines 387, 390, 391 and 401. It will be corrected.

18) Page 21 Line 406 – remove “a” to read “: : :.of good measurements: : :.” I don’t get the comparison to synthetic data. You site the advantages of using synthetic data are you know where the noise is, yet you don’t have plots showing a comparison to the known noise is?
A:

- Page 21 Line 406. It will be corrected. The position of the noise for an individual scan is shown in Figure 3 (c).

19) Page 27 Line 483 – replace “This” with “These” to read “These possible deviations: : :.”
A:

- Page 27 Line 483. It will be corrected.

Anonymous Referee #2

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.

General Comments

1) 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.
A: The procedural noise implemented here aims to generate V_LOS values smoother than the ones observed at very low CNR, and closer to "reasonable" V_LOS. Figure 1 below shows the distributions of synthetic, contaminated V_LOS values and real V_LOS data with CNR values below -32dB. From this Figure it is possible to see that the synthetic noise generates V_los closer to reliable values and thinner tails. The consequence of this is a more subtle contamination, which is harder to detect by the filters presented in this work. Additionally, the principle of coherent noise is to generate areas of contamination that are smoother in space, which also makes more difficult differentiate contaminated observations and clean data via the distance generated by ΔV_LOS and for DBSCAN, and a fixed threshold for the median-like filter. In summary, the intention of this implementation is not to recreate real noise (its nature is relatively unknown), but to test the filter in harder conditions than real situations.

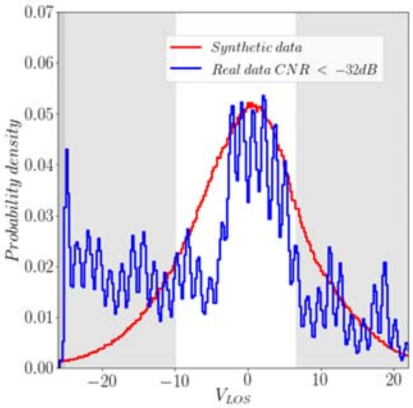

Figure 1: Pdf of V_LOS for contaminated synthetic data (one mean wind speed direction) and real data from the Balconies experiment at 200 m.a.g.l.

2) The math of the methods is not presented clearly in equations, especially regarding the filters.

A: The iterative operation of DBSCAN on discrete data is non-linear and is defined algorithmically. To the best of my knowledge, there are not references of transfer functions or reduced mathematical expressions of its frequency response for instance. Regarding the median-like filter, as mentioned in the paper, its most obvious parallel is the median filter used in image processing. As DBSCAN, this filter is non-linear, and it lacks a defininition in frequency domain as a transfer function. The development of theoretical expressions in this sense is out of the scope of this work. ### This is stated in the corrected version of the paper.

3) The work is not referencing important work in the field of lidar simulation and data filtering adequately.
   A: The suggested references will be checked and included accordingly.
4) 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.
   A: The references mentioned were considered and the description of the simulator will be rewritten in a more concise way.
5) 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.
   A: The coherent noise implemented here is not linear and is defined, as a parallel to DBSCAN, algorithmically. There is not a clear transfer function that defines it a priori and it can be better described by the V_LOS distribution after contamination. Figure 1 will be included in the corrected version.
6) 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.
   A: Section will be reworded in a more concise way.
7) 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?
   A: As mentioned earlier, the non-linear median-filter is defined algorithmically, not via equations. The filter is applied on single scans (it is a filter that operates spatially), and this is clarified in the corrected text. The filter closely related to the one described by Menke (2019), since it uses a moving window in the laser beam direction, the first stage. Nevertheless It does not applies a global filtering stage, which is replaced by a second moving window in the azimuth direction. The paragraph was reviewed and corrected to make it more clear.
8) 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.

A: As mentioned in 2), there is not a mathematical expression in the form of a transfer function for DBSCAN, since it is defined algorithmically. The connection between lidar parameters, or features, and DBSCAN is the definition of the observational space, where all (Euclidian) distances are calculated. This is better explained in the corrected text, as well as the k-distance function, or the function of the distance of each point to its k-nearest neighbor.

9) 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: ▯ recov) and the false positive rate (i.e. the percentage of data points that are considered good observations although they are contaminated by noise).

A: The false positive rate (false negative in the work, positive is noise detection) is equivalent to the fraction of noise detected, but it does not consider information of the fraction of contaminated observations in the scan. High recovery rates with low false positive rate (negative) might be only a low fraction of noise. This is the reason to include also a metric that takes into account the noise fraction.

10) 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.

A: I agree with this comment. The result of one of them will be only mentioned for a fair comparison in the corrected text.

11) Section 6.1: I think 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.

A: Since the results of synthetic scans are presented here, discussion on V_LOS threshold is better pictured fter the filters are applied. The discussion will be initiated in 4.2, and a figure comparing the two filters will be included in the corrected text.

12) 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.

A: Data recovery is very important in this work indeed, since the main motivation to explore a different filtering technique is to increase the amount of data available, which can be very poor when we use the most conservative CNR threshold. However, as the referee suggest, this is worthless if data quality is bad. This is the reason to complement the performance assessment with the metrics in 5.2. Figure 16 will be modified to consider this suggestion.

13) 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.

A: A change in the section title will be made to clarify its intention.

14) Section 7.3 and 8 could probably be combined.

A: Even though section 7.3 give some final remarks it still discuss on computational performance of the clustering filter and possible imporvements. The section will be revised to make this more clear.

Specific comments

1) p.1, l.1: simultaneous multi-point observations are possible with masts if multiple sonics are installed.
A: Indeed this is possible, at a high cost though, and it is not very common for wind resource assessment for instance. That is the meaning of the sentence.
2) p.1, l.2: write "lower" instead of "reduced"
A: It will be corrected in the text.
3) p.1, l.4: "reduced data recovery" compared to what? I am also not sure if "data recovery is the proper term.
A: The reduced data recovery is compared to the total amount of data available.
4) p.1., l.6: "...spatial position, and VLOS 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.
A: It will be rephrased for clarity.
5) p.1,l.13: "its adoption" - "their acceptance"!?
A: It will be corrected
6) 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.
A: It will be modified for clarity. The intention is to show values used in the reference cited.
7) p.2.l.37: typo "approaches"
A: It will be corrected.
8) p.2,l.56: "DBSCAN" acronym should be explained here.
A: It will be corrected
9) p.3,l.80: Why are the scanning patterns coherent?
A: It is intented to mean meaningful
10) p.5,l.105: The term "numerical lidar" is very unusual and irritating. I would recommend "lidar simulator" or "virtual lidar".
A: The term was already used in Meyer (2017) but the suggestion will be considered.
11) p.5,l.112: What does "coarse" mean here? Numbers should be given.
A: Coarser means a grid of range gates and beam spacing that is actually much coarser than the sampling spacing frequency of the lidar, and represents the spatial and time averaging of the instrument. The numbers of this are in table 3. The term will be clarified in the corrected text.
12) p.6,Eq.2: The variable names are somewhat confusing, because what is here ▯p is ▯R in the references of Smalikho and Banakh and ▯p in the references is rp here.
A: The intention was not to use the same notation as Smalikho and Banakh but Meyer (2017)
13) p.6,l.129: "corresponding range gate center"!
A: it will be corrected
14) p.6,l.130: "range gate length" is not very specific. If you give the explanation of rp from FWHM, you could also give the explanation of ▯p from the time window of the FFT.
A: Since this terms are well known, their definition was done in simpler terms.
15) p.7,l.149: typo "radial"
A: It will be corrected
16) p.7,l.154: referencing Figure 10 which is introduced much later, is bad style.

A: It will be corrected

17) p.7,l.158: type "in"
    A: It will be corrected
18) p.8,l.177: again, a figure (Figure 5) is referenced before its introduction.
    A: It will be corrected
19) p.9,l.180: "The fraction of beams contaminated at each band..."
    A: It will be corrected
20) p.9,l.183: typo "from".
    A: It will be corrected
21) 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.
    A: This will be removed in the corrected version.
22) p.10,l.215f: put citation Huang et al in parantheses.
    A: It will be corrected
23) p.13,l.251: the m in "m-dimensional" is not explained.
    A: It will be explained
24) p.14,l.285: How does dk(n) look like for the lidar signal problem?
    A: It is shown in figure 9. This figure will be probably removed in the final version.
25) p.15,l.298: typo: "noisy"
    A: It will be corrected
26) p.15,l.298: Figure 9 is referenced before introduction.
27) A: It will be corrected
28) p.15,l.302: Equation 6 is referenced before it appears. Please introduce it before.
    A: It will be corrected
29) Figure 8b) seems to be moreless the same as Figure 5b.
    A: They come from bad and good scans respectively. It will be probably removed in the corrected version.
30) p.17,l.333: "PDF" should be in capital letters as an abbreviation.
    A: It will be corrected
31) p.18,l.344: Something is wrong with the grammer in this sentence.
    A: It will be corrected, an “are” was missing.
32) p.18,l.345: What is the value of that is used in this study?.
    A: The value of alpha is 0.05
33) p.18,l.345: Again, grammar.
    A: It will be corrected
34) p.18,l.345f: The numbers about the amount of data that was analyzed should be given in Section 2.
35) A: It will be corrected
36) p.19,l.372: remove one "then".
37) A: It will be corrected
38) p.20,l.386: typo: "account"
    A: It will be corrected
39) Figure 11: I think this figure is not neccessary. If it is still shown, labels have to be larger.
    A: I agree and it will be removed in the corrected version.
40) Figure 14: typo, should be "phase 2"
41) A: It will be corrected
42) Figure 15: Why is no upper threshold for the CNR filter applied, which would remove the wind turbine hard target from the recovered data?

A: It is applied indeed, this will be written more clear.

43) Figure 16: I think this plot is not neccessary.
A: This comment will be considered.

44) 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?
A: It refers mostly to extreme values when compared to more reliable CNR data distribution.

45) p.23,l.443: Metrics are introduced in Sect. 5.1.
A: It will be corrected

46) p.23,l.443f: Again, quality is undefined.
A: It will be better explained

47) p.27,l.502: typo: "from"
A: It will be corrected

List of changes

1) Changes (rewording) in the introduction and data description in section 2.
2) Changes in Figures suggested by reviewers. Figures 3 (now 2), 4 (now 3), 5 (now 4) and 16 (now 13) were modified.
3) Figures 7, 8 and 9 are replaced by Figure 6.
4) Figure 11 is now A1 in appendix A.
5) Changes in Sections 3 and 4 regarding methodology description (Lidar simulator in 3.2, noise generation 3.3 and filters description in 4.1, 4.2 and 4.2) accounts for comments from reviewers.
6) Section 6.1 was modified and results in Figure 8, suggested by Anonymous Reviewer # 2, included.
7) Discussion in sections 7.1, 7.2 and 7.3 were merged in only one section.
8) Part of 7.3 was moved to 8.
9) Part of sections 4.2 and 6.1 were moved to Appendix A.

Filtering of pulsed lidars data using spatial information and a clustering algorithm

Leonardo Alcayaga[1]

[1]DTU Wind Energy, Frederiksborgvej 399, 4000 Roskilde, Denmark

Correspondence: Leonardo Alcayaga (lalc@dtu.dk)

Abstract. Wind lidars present advantages over meteorological masts, including simultaneous multi-point observations, flexibility in measuring geometry, and reduced installation cost; but wind lidars come with the ‘cost’ of increased complexity in terms of data quality and analysis. Carrier-to-noise ratio (CNR) has been the metric most commonly-used to recover reliable observations from lidar measurements, but with severely reduced data recovery. In this work we apply a clustering technique to identify unreliable measurements from pulsed lidars scanning a horizontal plane, taking advantage of all data available from the lidars—not only CNR, but also line-of-sight wind speed (V_{LOS}), spatial position, and V_{LOS} smoothness. The performance of this data filtering technique is evaluated in terms of data recovery and data quality, against both a median-like filter and a pure CNR-threshold filter. The results show that the clustering filter is capable of recovering more reliable data in noisy regions of the scans, increasing the data recovery up to 38% and reducing by at least two thirds the acceptance of unreliable measurements, relative to the commonly used CNR-threshold. Along with this, the need for user intervention in the setup of data filtering is reduced considerably, which is a step towards a more automated and robust filter.

1 Introduction

Long range scanning wind lidars are useful tools, and ~~its~~ their adoption has grown rapidly in recent years in wind energy applications (Vasiljevic et al., 2016). ~~Their capability to measure the evolution in time and space of atmospheric boundary layer wind fields in large spatial domains(which can reach a size comparable to a wind farm), their increasing accuracy and low~~ Scanning wind lidars can measure time evolution and spatial characteristics of wind fields over large domains, at a lower cost of installation ~~are important advantages over~~ than meteorological masts. ~~One disadvantage of these devices with respect to traditional wind speed measurement techniques is the influence that~~ Nevertheless, atmospheric conditions and instrument noise ~~have~~ can have an important impact on the data quality. For long-range scanning lidars ~~,~~ this becomes an important issue due to the lack of ~~references to identify reliable observations~~additional instruments placed over the measurement area that would be useful to compare data quality, since noise can contaminate large portions of the scanning domain. The most commonly used criteria to retrieve reliable observations is a threshold on values of the Carrier-to-noise ratio, CNR, ~~which, depending on the~~ threshold that will depend on site conditions, experimental setup and ~~instrument manufacturer, can take values between -29 dB, -20 dBand -8 dB, 0 dBas lower and upper bounds, respectively~~the instrument manufacturer (Gryning et al., 2016; Gryning and Floors, 2019). ~~This criteria results~~ Despite CNR threshold retrieve quality observations, its application might result in large amounts of ~~data rejected unnecessarily~~ good data rejected in regions far from the instrument, ~~due to the nature of CNR which decreases~~ where CNR has decreased rapidly with distance. To cope with this issue Meyer Forsting and Troldborg (2016) and Vasiljević et al. (2017) have proposed filters based on the smoothness and continuity of the wind field. Such filters work by detecting discrete or anomalous steps ~~in V_{LOS} which present a difference between of~~ (above a certain threshold, predefined by the user) in line-of-sight wind speed, V_{LOS}, ~~and~~ compared to its local (moving) median~~above a certain threshold, predefined by the user~~. Beck and Kühn (2017) first and Karagali et al. (2018) in an adapted version, follow a different approach (here called KDE filter, from Kernel Density Estimate) based on the statistical self-similarity of the data, which, in simple terms, ~~assumes~~ means that reliable observations are alike and will be located close together in the observational space. The probability density distribution of observations (estimated via KDE) in a dynamically normalized $V_{LOS} -$ CNR space shows that measurements likely to be valid are located in a high data density region. Observations sparsely distributed beyond a boundary defined by a threshold in the acceptance ratio, or the ratio between the probability density of any observation and the maximum probability density over the whole set of measurements, are finally identified as noise. Both approaches need the definition of one or more thresholds and a window size, either in time for the KDE filter, or in space for the wind field smoothness approach. These parameters are dependent on different characteristics of the data, like the lidar scanning pattern for instance.

Both approaches miss important and complementary information, either neglecting the ~~quality of acquired data (quantified in terms of~~ stregth of the signal back-scattering (quantified by CNR) or the spatial distribution and smoothness of the wind field. Moreover, in both ~~apporaches~~ approaches the position of observations is not taken into account, information that can shed light on areas permanently showing anomalous values of V_{LOS} or CNR, like ~~e.g.~~ hard targets. ~~Complementing~~ Including all these features within the smoothness approach is difficult, since CNR is not a smooth field like V_{LOS}~~and including them~~. Moreover, considering smoothness and position in the KDE filter ~~increase the computational time substantially, since the basis to define high density regions and an acceptance ratio is a kernel density estimate~~ results in a computationally costly kernel density estimation, if we look for an optimal bandwidth parameter in a higher dimensional space~~(3 or more features including spatial position and smoothness), which is computationally intensive due to the estimation of a bandwidth parameter and the definition of a high density region (hyper-volume) with a good resolution~~, with a fine resolution of the kernel density estimate.

Data self similarity – over any scale in the case of fractals or a range of them in real situations (Mandelbrot, 1983) – is closely related to clustering techniques (Backer, 1995), which ~~are capable of~~ can classify large data sets with many different features at a relatively low computational cost. ~~For instance, the KDE filter~~ The KDE filter approach shares some characteristics with the popular k-$means$ clustering algorithm~~(first presented by MacQueen (1967))~~ MacQueen (1967), since they define one (or several for $k-means$) specific group of data belonging to an unique category (or cluster) ~~which~~ whose size and location on the observational space will depend on data density or, more specifically, on a kernel density estimation. The main difference between these two algorithms is the way they treat sparse data points that fall in low density regions. ~~In $k-means$, sparse points are assigned~~ Unlike the KDE filter, which rejects noise via the acceptance ratio, $k-means$ assigns sparse points to the cluster with the nearest center, no matter if they are outliers or present unlikely values from a physical point of view. ~~The KDE filter solves this problem introducing an acceptance ratio, which corresponds to a threshold on the probability density of data points that are to be accepted as cluster members. This threshold must be defined a priori when used on unfiltered data.~~

~~Additionally, $k-means$ needs to define the number of clusters present in the data beforehand, unlike the KDE filter, which defines always an unique cluster of valid data points, centered at the origin of the scaled and normalized observational space.~~

~~The~~

The Density Based Spatial Clustering for Applications with Noise algorithm, or DBSCAN~~clustering technique (Ester et al., 1996; Pedrego~~ (Ester et al., 1996; Pedregosa et al., 2011), introduced in Section 4.3, presents several advantages over k-$means$ in detecting clusters in a higher dimensional space: ~~1)~~ it introduces the notion of noise/sparsely distributed observations~~and 2)~~ , it does not need prior knowledge of the number of clusters in the data and it is capable of ~~identify~~ identifying clusters of arbitrary shape. To the best of our knowledge, this is the first time that this type of clustering algorithm is applied to identify not reliable observations from pulsed lidars. This approach, which can be understood as a natural extension of the KDE filter, is compared to the smoothness based filter on two types of data: synthetic wind fields data as a controlled test case, and real data.

This paper is organized as follows: Section 2 describes the real data used to test the different filtering approaches, and Section 3 presents the synthetic data used during a controlled test as well as the methodology to obtain it. Section 4 then gives a description of the different filters applied in this study to both data sets, to continue with the definition of the performance tests in Section 5. In Section 6 the ~~performace~~ performance tests are presented along with a discussion on their validity and quality. Section 7 discuses the quality of the methodology behind the tests and the advantages and disadvantages of the proposed approach. Section 8 presents the conclusions of this study.

2 Real data: Østerild Balconies experiment

The filtering techniques presented here were tested on lidar measurements made at the Østerild Test Centre located in northern Jutland, Denmark, see Figure 1. ~~The aim of this experimentwas to characterize the horizontal flow field~~ Known as the Østerild Balconies experiment (Mann et al., 2017; Karagali et al., 2018; Simon and Vasiljevic, 2018), this measuring campaign aimed to characterize horizontal flow patterns above a flat, heterogeneous forested landscape at two heights relevant for wind energy applications~~. Known as the Østerild Balconies experiment (Mann et al., 2017; Karagali et al., 2018; Simon and Vasiljevic, 2018), the wind speed data covers a large horizontal area (~~, covering an area of around 50 km^2~~), with the possibility of characterizing flow patterns in~~ , and a wide range of scales, both in time and space. ~~However, these advantages come with increased complexity on data reliability. A larger measurement area is affected by local terrain and atmospheric conditions, like clouds or large hard targets. Moreover, at this scale lidars reach their measuring limitations, since the back-scattering from aerosols decrease rapidly with distance (Cariou, 2015).~~

~~The Balconies experiment consists~~

The experiment consist of two measuring phases (see Table 1) with two long-range WindScanners performing Plan Position Indicator (PPI) scanning patterns, aligned in the North-South axis and installed at 50 m a.g.l. during phase 1 and 200 m a.g.l. in phase 2. WindScanners (Vasiljevic et al., 2016) consist of two or more spatially separated lidars which are synchronized to perform coherent scanning patterns, allowing the retrieval of two or three dimensional velocity vectors at ~~diffeent~~ different points in space. These experiments were conducted between April and August of 2016 (Simon and Vasiljevic, 2018). In each

Table 1. Characteristics of the Balconies experiment, from Karagali et al. (2018). The scans are not instantaneous neither totally synchronous, with a horizontal sweep speed of 2°/s in the azimuth direction in a range of 90°, with a total time of 45 s per scan.

Phase	**Measurement start**	**Measurement end**
50 m a.g.l. (1)	2016-04-12 12:45:41	2016-06-17 12:48:01
200 m a.g.l. (2)	2016-06-29 13:35:56	2016-08-12 09:09:55
Scanner	**Location coordinates, [m]**	**Scanning pattern, west**
Southern lidar	492768.8 (East) 6322832.3 (North)	344°-256°, 2° steps
Northern lidar	492768.7 (East) 6327082.4 (North)	196°-284°, 2° steps

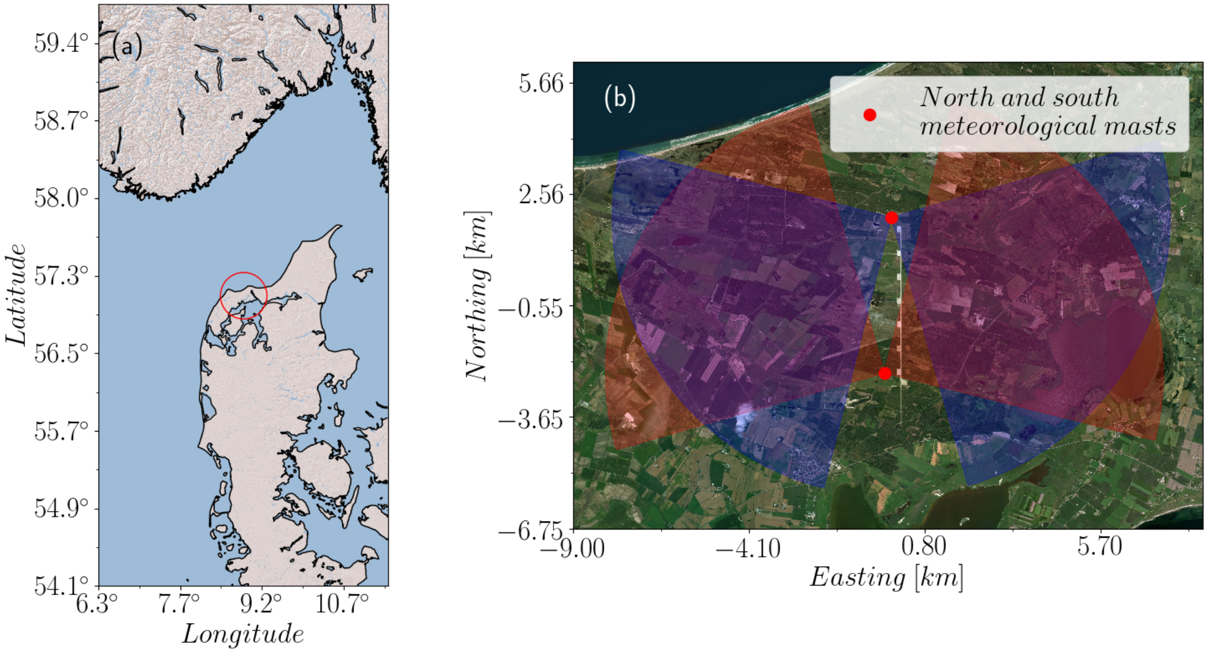

Figure 1. (a) Location of the Østerild Turbine Test Center, place of the Balconies experiments, northern Jutland, Denmark (copyright 2009 Esri). (b) Detail of the test center site, with the location of the meteorological masts where north (blue) and south (red) WindScanners were installed. During the measurement campaign the PPI scans pointed both west in some periods and both east in other (copyright 2017 DigitalGlobe, Inc.).

phase, the northern and southern lidars scanned in the West and East direction relative to the corresponding meteorological masts, where they were installed. The data used in this study originated from both phases of the experiment, with PPIs pointing to the west. For more details about the experiment, lidars and terrain characteristics see (Karagali et al., 2018; Vasiljevic et al., 2016; Simon and Vasiljevic, 2018).

This dataset is well suited to test different data filtering techniques. A large measurement area will be affected by local terrain and atmospheric conditions, like clouds or large hard targets. Moreover, at this scale lidars reach their measuring limitations, since the back-scattering from aerosols decrease rapidly with distance (Cariou, 2015).

3 Synthetic data

Assessing and comparing the performance of filters is challenging with no reference available to verify that rejected or accepted observations are ~~truly outliers or simply wrong~~ reliable or bad observations. This is especially difficult for long-range scanning lidars, since their measurements cover large areas and, due to spatial variability, a valid reference would need several secondary anemometers scattered over the scanning area. Testing filters on a controlled and synthetic data set, contaminated with a well defined noise, presents an option to deal with this problem. In this study, the filters presented in Section 4 are tested on individual scans sampled from synthetic wind fields generated using the Mann turbulence spectral tensor model (Mann, 1994), and contaminated with procedural noise (Perlin, 2001).

3.1 Synthetic wind fields generation

Synthetic PPI scans are sampled by a ~~numerical lidar~~ lidar simulator from synthetic wind fields generated via the Mann-model (Mann, 1998) in a horizontal, two-dimensional square domain of 2048 x 2048 grid points, with dimensions 9200 m x 7000 m. The generated turbulence fields are the result of input parameters of the of turbulence spectral tensor model~~: –~~, namely, length-scale, L, turbulence energy dissipation $\alpha\epsilon^{2/3}$, and anisotropy, Γ. The fields generated correspond to wind speed fluctuations, ~~which are subsequently added later to a mean flow field to generate the resulting synthetic wind speed fields. An~~ to which the desired average wind speed mean is subsequently added. Depending on the initial random seed ~~generate~~ used, different wind field realizations with the exact same turbulence statistics can be generated. For details on wind field generation using the the Mann-model, refer to Mann (1998). Table 2 shows the range of values used for the generation of two-dimensional wind fields. Large values of $\alpha\epsilon^{2/3}$ or ~~smaller~~ small scale turbulence for instance, ~~make~~ mean that sudden spatial changes in wind speed are more likely, which increase the false identification of outliers. Mean wind direction, turbulence anisotropy and length scale will also affect the sampling due to the lidars ~~sampling~~ measuring characteristics.

3.2 ~~Sampling with a numerical lidar~~Lidar simulator

~~The numerical model of a long-range pulsed lidar attemps to mimic what the real instrument does obtaining, for instance, the data described in Section 2. Even though its implementation is very crude andsimplified, it allows to generate sampled scans with reasonable spatial smoothness and complexity, but it is, however, not able to reproduce either noise nor CNR values. Since the interest of this study is to test filters over a large number of scans, its simplicity allows a quick sampling over high resolution synthetic wind fields , thus reducing computational time~~

Lidar simulators has been presented previously by Stawiarski et al. (2013) and Meyer Forsting and Troldborg (2016). They sample V_{LOS} values from wind fields generated via Large Eddy Simulations (LES), mimicking the operational principle of

Table 2. Synthetic wind field characteristics and parameters.

Parameter	**Values**
L, m	62, 125, 250, 500, 750, 1000
$\alpha\epsilon^{2/3}$, m$^{4/3}$s^{-2}	0.025, 0.05, 0.075
Γ	0, 1, 2, 2.5, 3.5
Number of seeds used	10
Mean wind speed, U m/s	15
Mean wind speed direction range, degrees	90 to 270
Total number of scans generated	4305

lidars by proper time and spatial (probe volume) averaging of the background wind field. The lidar simulator presented here follows the same principles, this time sampling from synthetic wind fields generated via Mann-model. ~~The sequence followed by the model in sampling from the high resolution wind fields is specified below (see also Figure ?? for more details): A coarse two dimensional mesh in polar coordinates (r, θ) is generated, with radial (range gates) and azimuth ranges and resolution described in Table 3. A set of non overlapping nested meshes are then constructed locally, centered in each point in an upper level, coarser mesh . Each nested mesh has 21 x 51 grid points in the radial and azimuth directions, respectively. The resolution in the radial direction of the nested mesh guarantees at least 1 synthetic observation within grid elements located in the two beam range gates closest to the lidar(the closest range gate in this case is 150m from the lidar) , since, even though the resolution of~~

The simulator receives scanning pattern characteristics as input (beam range, range gate step, azimuth angles range and azimuth angle steps) to generate a primary mesh with the sampling positions on top of background wind field. Following the measuring principle of the lidar, the V_{LOS} observed at each position in this mesh will represent averages of a continuous along each range gate step (due to probe volume averaging) and an average of many azimuth positions within the azimuth step, due to the almost continuous sweep of the ~~wind field Cartesian grid is high, the size of its elements is comparable with the nested ones close to the numerical lidar.~~ lidar's beam. The simulator mimics this generating a secondary, refined mesh with N_r points in each range gate and N_ϕ beams within each azimuth step. The ~~number of elements in the azimuth direction allows a low grid aspect ratio at the end of the beam, which is 7km from the lidar. The streamwise and lateral wind speed components from the synthetic wind field~~ background wind field components, U and V, are ~~linearly interpolated into the nested fine mesh and the radial component (~~then interpolated on this secondary mesh and projected on each refined beam to obtain V_{LOS} ~~) calculated using~~ using equation (1), with θ being the ~~azimuth angle of the local beam~~corresponding beam azimuth angle.

$$V_{LOS} = \cos(\theta)U + \sin(\theta)V \tag{1}$$

~~The numerical lidar mimics the volume averaging in a real lidar by averaging in the radial direction each of the 51 beamsegments in the local grid, using~~ The final step is the spatial (probe volume) averaging, and the azimuth (sweeping) averaging around each position in the primary mesh. Spatial averaging is done applying a weighting function on all V_{LOS} ~~values weighted with~~

~~w defined by~~

$$w = \frac{1}{2\Delta p}\left\{\mathrm{Erf}\left[\frac{(r-F)+\Delta p/2}{r_p}\right] - \mathrm{Erf}\left[\frac{(r-F)-\Delta p/2}{r_p}\right]\right\}$$

along each refined beam. The weighting function used here is defined in equation (2), as in Banakh and Smalikho (1997) and Smalikho and Banakh (2013). ~~Here F is the distance from~~ This function will assign weights to each point in the ~~beam to the corresponding range gate , Δl is the lidar beam's~~ refined beam according to its distance to the range gate position in the primary mesh, F, and the instrument probe volume parameters, namely, range gate length, Δp, and full width at half maximum~~(provided by the lidar's manufacturer), Δp is the range gate length~~ , Δl (cf. Table 3)~~,~~

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}}\int\limits_0^x \exp(-t^2)dt$$

~~denotes~~ . Here, Erf(x) is the error function, and

$$r_p = \frac{\Delta l}{2\sqrt{\ln(2)}}$$

~~is~~ r_p the beam width contribution to the volume averaging. ~~This kernel is not truncated, having influence of back-scatter from more distant positions on the local estimation, thus smoothing the V_{LOS} final field. Even though w decays rapidly with distance for the numerical lidar setup used here, there is still some influence from points located around distant range gates.~~

$$w = \frac{1}{2\Delta p}\left\{\mathrm{Erf}\left[\frac{(r-F)+\Delta p/2}{r_p}\right] - \mathrm{Erf}\left[\frac{(r-F)-\Delta p/2}{r_p}\right]\right\} \;;\quad r_p = \frac{\Delta l}{2\sqrt{\ln(2)}} \tag{2}$$

~~The result of this beam averaging is one radial velocity,~~ The azimuth averaging is the arithmetic mean of the N_ϕ values of V_{LOS} ~~, per range gate , 51 values in total per nested grid, along one arc in the azimuth direction. Pulsed lidars accumulate~~ at each range gate after spatial averaging. It represents the accumulation information of the back-scattered signal spectra as they sweep an azimuth sector ~~(2 degrees in our case)~~ before estimation of the spectral peak and V_{LOS}. ~~This continuous sweep is modeled here by the 51 discrete values of V_{LOS} along the azimuth arc and their non-weighted average. This step generate radial wind speeds in each point of the initial coarse polar grid, which constitutes the synthetic scan.~~

~~The local fine polar mesh and averaging in the azimuth direction here play the role of the rapid sweeping of the laser beam, taking into account that the measured V_{LOS} will be the result of the accumulation of information on the back-scattering spectra within the 2° azimuth steps, as well as the radial/line averaging around each range gate. This model is, of course, very simplistic and does not contain information about back-scattering spectra nor Doppler effect. Moreover, it does not mimic the non-instantaneous nature of the real scans, but rather assumes that all beams cross the measuring domain simultaneously, recording radiual wind speeds at the same time. Nevertheless, it will generate scans with radial speeds projected from a realistic wind field and can be contaminated with noise (presented in Section 3.3) under controlled conditions, which is the final goal of this test.~~

Table 3. The characteristics of the ~~numerical~~ lidar simulator and real long-range lidar (Karagali et al., 2018; Vasiljevic et al., 2016) ~~long-range lidars~~ used for the controlled test of the filters.

	~~Numerical~~Simulator	**Real**
Azimuth range	256° - 344°	256° - 344°
Azimuth step	2°	2°
Beam length	7000 [m]	7000 [m]
Range gate length, Δp	35 [m]	35 [m]
~~Time~~ Full width at half maximum, Δl	75 [m]	75 [m]
Sweeping time per scan	Instantaneous	45 [s]
~~Coarse polar grid size (radial-azimuth~~Primary mesh size (radial x azimuth)	45 x 198	-
~~Nested (local) polar grid size (radial-azimuth~~Secondary mesh size at each range gate (N_r x N_ϕ)	21 x 51	-
Total secondary mesh size (N_r x N_ϕ)	21 x 51	-

~~The numerical lidar mimics the accumulation of information in the spectra of the back-scattered signal (within each 2° step) as the average of 51 discrete beams. The radial wind speed in each of the 51 local range gates is the result of a weighted contribution of 21 points along each local beam segment, which spans from the previous to the nest range gate. The final V_{LOS} will be the average of these 51 values. Here the weight, w, for beam averaging, originates from equation .~~

3.3 Synthetic noise generation

The most simple noise that can be used to contaminate synthetic scans is sparse, uniformly distributed outliers~~, taking extreme values from the tails of non reliable observations (see Figure 7), and always within the detectable range of ± 35 m/s, characteristic of the pulsed lidar described in Section 2~~. This noise, also known as salt and pepper noise, is easily detected and eliminated by median-like filters, when extreme discrete steps affect the smoothness of an image (Huang et al., 1979; Burger and Burge, 2008). Nevertheless, ~~what one can see from~~ noise in real scans ~~are~~ comes as regions of anomalously high and/or low V_{LOS} ~~values (see Figure 3 (a)), and , depending on their relative size to the moving window of a median-like filter (more details en Section 4), they could~~ and they can pass through the filter undetected~~and unchanged. A more realistic noise(and also a more difficult to detect, goal of the test), is the procedural noise generated via the simplex noise algorithm,~~ . Procedural noise, introduced by (Perlin, 2001) ~~originally to recreate more natural~~ to recreate synthetic textures on surfaces for computer graphics applications~~. This type of noise is an option to generate more natural noisy regions with smoother transitions between large V_{LOS} outliers. The basic principle behind this algorithm can be roughly summarized as follows:~~ , creates regions of coherent noise that resembles better the spatial distribution of scanning lidars measurements. For the two-dimensional case, the procedural noise function $N(x,y)$ maps two-dimensional coordinates, (x,y), onto the range $[-1,1]$ as follows,

- A ~~relatively coarse two dimensional grid of pseudo random unit gradients is generated~~two-dimensional grid of m by n elements is generated, and a pseudo-random, two-dimensional unit gradient, $\mathbf{g}_{ij} = (g_x, g_y)$, is assigned to each grid point

(x_i, y_j). The pseudo-randomness ~~is established as follows: a list of permutations (indexes permuted) with~~ rises from the fact that $\mathbf{g}_{ij}$ are picked from a pre-computed list of gradients with length $l << m \times n$. We select values from this list using the index permutation grid $p_{ij} \in \{0, ..., l\}$ also with $m \times n$ elements. Then, $\mathbf{g}_{ij}$ will correspond to the gradient in the position p_{ij} of the ~~same number of elements as the grid will sample gradients from a list of limited length (16 unit gradient elements in our case).~~ pre-computed list. Elements p_{ij} are shuffled for each realization.

- ~~A set of points p is arbitrarily distributed within the domain defined by the grid.~~
- ~~To estimate the noise level of each point p, the contribution from each of the nearest pseudo-random gradients is calculated as the dot product between the corresponding gradient and the distance vector , $\mathbf{d}$, from the gradient position to p. Finally, each contribution is added directly after weighting by the inverse of the magnitude of $\mathbf{d}$, and scaled to obtain values within the range -1, 1.~~
- ~~As one could suspect, each p matches one of the positions to contaminate in the synthetic scans.~~ For each grid point (x_i, y_j) enclosing (x, y), a distance vector $\mathbf{d}_{i,j} = (x - x_i, y - y_i)$ is generated.
- Finally, the noise function is the sum of dot products, $N(x,y) = \sum_q w_q (\mathbf{g}^q_{ij} \cdot \mathbf{d}^q_{i,j})$, for q grid points surrounding (x,y). Weights w_q correspond to $w_q = C \frac{1}{\|\mathbf{d}^q_{i,j}\|}$, and C a normalization constant to ensure that $N(x,y) \in [-1, 1]$.

~~When clouds, rain or atmospheric conditions affecting the concentration of aerosols in the air enter the measuring domain, what we see are regions with anomalous line-of-sight wind speeds and low CNR values. One of the advantages of procedural noise is the generation of areas with noise, with the same size of the scan domain for instance, either in polar or Cartesian coordinates, via an array of points p distributed over the synthetic measuring domain. In this test, the distribution of points p over the scanning area aims to follow the decay in the~~ The function $N(x,y)$ allows the generation of noisy regions, than can be distributed according to back-scatter ~~intensity with distance(See Figure 4). We define three bands per scan,~~ decay with distance. Three bands centered at 50% , 70% and 90% of the total beam length ~~,~~ (and spanning over the entire azimuth range~~and with a width of~~) have an increasing fraction of noise, contaminating the 30~~% of the beam length in the radial direction. Within each of these bands a set of uniformly distributed positions are selected. The fraction of beams contaminated at each band, as we depart from the lidar, are 30~~%, 60% and 90% ~~, respectively, since it is assumed that after a position of a particular beam is sampled, the remaining points in the radial direction are all contaminated. The increasing fraction of contaminated points tries to be similar to what is observed in real data, since the quality of the signal decrease as we depart form the lidar~~of the observation, respectively. The noise amplitude is ~~finally scaled by~~ 35 [m/s]~~; i.e. ,~~ the limit of the observable range ~~of V_{LOS}~~ for the instruments described in Section 2.

Figure 2 (~~a) and (b) show the~~ c) show one contaminated scan and its increasing contaminated area as we move along the beams. The same Figure shows the distribution of the noise generated by the ~~simplex~~ algorithm after scaling, and ~~its effects on the distribution of~~ the probability distribution of contaminated synthetic V_{LOS} ~~, respectively~~compared to real data with low values of CNR. The distribution of ~~contaminated observations~~ real data presents heavier tails than the ones generated, with higher probability of observing extreme values of V_{LOS}. Modeling real noise is difficult, ~~as expected, but also a high probability~~

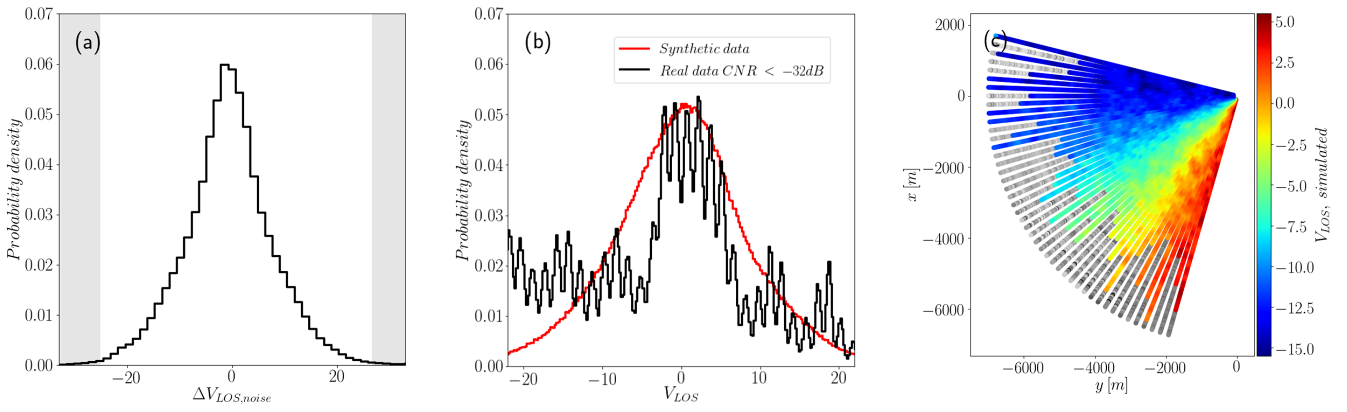

Figure 2. ~~Contamination with procedural~~ Procedural noise on synthetic scans.(a) Distribution of $\Delta V_{LOS,noise}$, the noise added~~to the original scan~~. Maximum values are within the observable range between ~~−35 and~~ [-35, 35] [m/s]. (b) Distribution of ~~non-contaminated~~ real V_{LOS} with low CNR values (black) and contaminated, synthetic $V_{LOS} + \Delta V_{LOS,noise}$ (red) ~~,~~ for ~~all range gates and cases. The distribution of V_{LOS} show peaks around ±15 m/s, which correspond to the main~~ a mean wind ~~directions simulated~~direction facing the scan. (c) Individual scan showing ~~a realization of~~ the ~~spatial distribution of radial wind speed, V_{LOS}, with~~ increasing fraction of added noise ~~$\Delta V_{LOS,noise}$~~ (grey) with distance.

~~for V_{LOS} values within a range far from the shadowed area of outliers. Therefore, contaminated points are more difficultto spot by filters that do not consider information about spatial smoothness. Figure 2 (c) shows one contaminated synthetic scan, where the noise fraction increases with distance. The noisy areas show relatively smooth transitions in the azimuth direction due to the grid of gradients is generated in polar coordinates, which makes the noise identification more difficult for median-like filters, which are good at detecting local discrete steps in V_{LOS}, but have difficulties to do it when the size of its moving window is comparable to the size of the contaminated area~~since the process that generates it depends on the measuring principle of the lidar and atmospheric conditions. The synthetic noise used here does not intend to be totally realistic, but more subtle and smoother than the one observed in real measurements, making the identification of contaminated points more difficult.

4 Filtering techniques ~~tested~~ applied on real and synthetic data

4.1 CNR threshold

~~Along with the line-of-sight wind speed, lidars give information about the intensity of the back-scattered signal via CNR , which can take values in a range from 15 to -50 dB, depending on the manufacturer of the instrument. Here, the low CNR values correspond to very poor signal back-scattering, due to atmospheric conditions (i.e. low concentrations of aerosols) and high ratios usually occur when the lidar laser beam hits a hard target~~CNR thresholds are well known and lidar manufacturers usually recommend values for rejection of signals with poor backscattering or hitting hard targets (Cariou, 2015). ~~Good and reliable~~

~~measurements will lay in between these too scenarios, and thresholds values have traditionally been used to filter out non reliable observations (Gryning et al., 2016). Observations with CNR values lying beyond upper and lower thresholds (which will depend on the lidar itself, commonly -8 and -24 dB, respectively) are rejected. One problem that arise with the use of this filtering criteria for long range lidars is that the CNR value worsen rapidly with distance, and observations which might be valid at spatial points that are distant from the instrument are rejected using this criteria. Figure 3 (b) shows a scatter plot of radial wind speed, V_{LOS}, against CNR, from 30 consecutive PPI scans from the Balconies experiment, with its typical comb-like shape. Below the lower threshold (in this case~~ However, the selection of an appropriate threshold for CNR that assures data quality and good data recovery is not easy. Figures 3 and 4 show data from a scan with noisy observations from CNR values below -27 dB~~) along with extreme~~. Both, extreme and limited values of V_{LOS}~~we also find line-of-sight speeds that are not far from the reliable range above the threshold. The effect of filtering out only observations below the threshold or accepting the the ones with values within the range of V_{LOS} above the threshold can be seen in Figure 3 (c)and (d). Low CNR values for reasonable line-of-sight speeds can be understood when the range distance is included in the picture, as in~~, show low CNR values in the distant region of the scan, and data loss results after the application of the CNR threshold (Figure 4 (b)). When a limit to V_{LOS} is applied instead, Figure 4 ~~: the decay in the back-scattering intensity with distance makes CNR values worse, but, apparently, the lidar is still able to retrieve some valid observations. The selection of an appropriate threshold for CNR is not clear, but a lower bound of -29 dB is recommended by Gryning and Floors (2019) before a wind resource assessment, based on lidar measurements, starts to be dependent on the CNR-threshold~~(c) show that the smoothness in V_{LOS} is lost in the lower part of the scan. A conservative threshold of -24 dB is used here, since the resulting V_{LOS} probability distribution show very little outliers and it can be used as a reference when the performance of the filters proposed are compared.

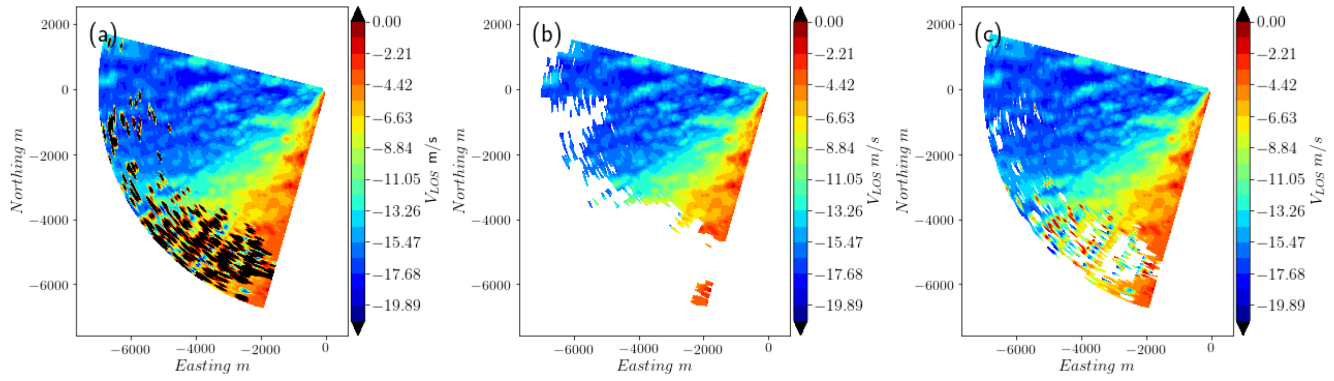

Figure 4. (a) ~~When range gate distance is included in~~ Un-filtered scan with V_{LOS} values outside the ~~picture~~range [-21,~~the shape of the distribution of reliable observations gets more complicated, and~~ 0] in Figure 3 (a~~decreasing CNR with distance is now obvious~~) in black. ~~Observations~~ (b) Filtered scan with ~~low~~ CNR ~~values but high probability density can be found at~~ > -27 and the ~~more distant region~~ resulting data loss in the uper part of the scan. (~~b~~c) ~~Threshold in the actual values of~~ V_{LOS} ~~(shadowed volumes) recovers distant~~within the [-21,~~good measurements~~0] range, ~~but still includes non reliable observations with very low probability density or KDE~~ showing anomalous values in the lower part of the scan.

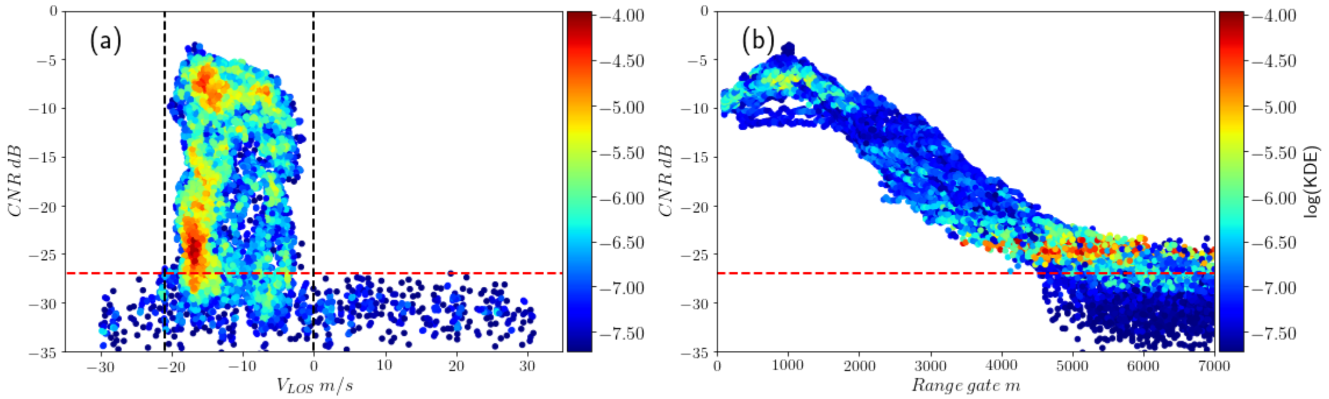

Figure 3. (a) ~~Scan from data measured in the Østerild Balconies Experiment with noisy observations in the far region. (b) Distribution of~~ CNR and V_{LOS} for ~~the same~~ one scan from Balconies experiment, including the probability density ~~estimated via~~ (KDE). Observations with ~~high~~ CNR~~values~~ >-27 dB (dashed red line) show a limited range of V_{LOS} (~~black~~ dashed black line). A ~~CNR threshold (dashed red line) of -27 places a~~ portion of ~~reliable~~ observations ~~that belong to a~~ with high probability density ~~region~~ remain in the rejection area. (~~e~~b) ~~Since points~~ CNR v/s distance for the same data. Observations with low CNR ~~are located~~ values and high probability density can be found in the ~~far~~ distant region of the scan~~, a large portion of the measured area is lost after filtering~~.~~(d) Values out of the range defined by the dashed black lines in (d) are rejected instead. The measured area gained still shows some noisy observations~~

4.2 Median-like filter

The ~~main output of scanning lidars is, in the present measuring campaign, a two-dimensional field of line-of-sight wind speeds. Interpreting this field as an image, a~~ median filter arise as a viable option for detecting erroneous measurements, since it is well known that this type of non-linear filter is suited to detect and filter noise that present distributions with large tails ~~Huang et al. (1979)~~. Here we use an adaptation of the traditional median filter used in the ~~image processing community: values~~ image-processing community, closely related to the three-stage filtering technique described in Menke et al. (2019): observations are not replaced by the local moving median but ~~labeled as reliable or non reliable according to whether their values are either under or above a threshold for $\Delta V_{LOS,\,threshold}$,~~ excluded if the absolute difference between ~~the value and its moving median. Another difference is the two-dimensional nature of the original median filter, which estimate the local median within a two dimensional moving window. Here,~~ their value and the ~~median-like filter does the same but in~~ local moving median is above a certain threshold, $\Delta V_{LOS,\,threshold}$. Unlike Huang et al. (1979), The two-dimensional moving window is replaced by a two one-dimensional window instances, the first in ~~the~~ line-of-sight or radial direction, r, and finally in the azimuth direction, θ, considering the polar coordinates of the scan. This simplification reduces the computation time ~~, which is the main advantage of this filter.~~ importantly.

The input parameters of this filter will be the size (or number of elements) of the moving ~~window~~ windows in the radial ~~direction,~~ and azimuth directions, n_r ~~, the size of the window in the azimuth direction, n_ϕ,~~ and ~~a threshold for the difference between the local radial wind speed value and its corresponding moving median,~~ n_ϕ respectively, and $\Delta V_{LOS,\,threshold}$. For fixed values of $\Delta V_{LOS,\,threshold}$, n_r and n_ϕ, the spatial structure of wind speed fluctuations will have an important effect on the recovery rate and noise detection of this filter. ~~Section 6.1 explores this relationship by means of a sensitivity analysis on the performance of this filter when applied over wind fields with different characteristics.~~

4.3 ~~Filtering using a clustering algorithm~~

~~Assuming no abrupt spatial changes in the different features measured, radial wind speed for instance, observations non affected by poor back-scattering or noise generated in the lidar itself, will fall in limited regions of the observational space, unlike contaminated observations, that will be scattered in wider region. This is noticeable in Figure 4, distinguishing two main groups: one presenting limited radial wind speeds and CNR values (between -20 and 0~~ A sensitivity analysis carried out using the metrics presented in section 5.1 on the synthetic data set, shows that the optimal set of parameters is n_r = 5, n_ϕ = 3 and $\Delta V_{LOS,\,threshold}$ = 2.33 m/s ~~and 0 and -35 dB, respectively)and another group, less dense and broader, with bad observations. Also noticeable is an overlapping region that contains observations from both groups. When only V_{LOS} and CNR are taken into account (Figure 3 (b)), the overlapping region is more diffuse than when including the range gate distance in the picture (Figure 4). The two groups are more distinct in the latter representation because their separation increases along with dimensionality as we consider more features of the data. These considerations inspire us to think the identification of valid observations as a process, which classifies groups or *clusters* of good/bad data, using all the information available: temporal~~

~~and spatial information, signal quality via CNR, V_{LOS} values and its smoothness.~~ (See Appendix A). This set is used both for artificial and real data. The filter does not include a time window, and it is applied on individual scans.

~~There exist many clustering algorithms with different characteristics and performance (Rui Xu and Wunsch, 2005; Xu and Tian, 2015), each of them suitable to the specific characteristics of the databeing analyzed. Partitions algorithms, like *k-means* (MacQueen, 1967) or *k-medoids* (Park and Jun, 2009) have been popular in the data mining community due to its low computational complexity. In these algorithms data points are separated in k different groups or clusters, with each observation belonging to the cluster with the closest mean/medoid, estimated iteratively by minimization of the within-cluster variance. Even though these algorithms introduce the notion of density in the data distribution, they have two drawbacks: both need prior knowledge of the number of clusters present in the data and, if this first problem is sorted, the algorithm assigns all data points to specific clusters, either good measurements or outliers, which is not desirable for our purpose. The Density Based Spatial Clustering for Applications with Noise algorithm, or DBSCAN (Ester et al., 1996), on the other hand, is a clustering technique specially designed to deal with large-scale data with spatial distribution. When compared to other algorithms, DBSCAN presents several advantages when applied as a filter to measurements from lidars: 1) it can manage large amounts of data with spatial distribution (its time complexity is $O(n\log n)$, with n the amount of~~

4.3 Filtering using a clustering algorithm

If we represent lidar observations as m-dimensional ~~points in the dataset), 2) it does not need prior knowledge of the number of clusters in the data, 3) it is a non-supervised algorithm (meaning that it does not require any fitting with previous training data), 4) it identifies clusters with arbitrary shapes and 5) it does introduce the notion of a noise set, which are data points that do not belong to any cluster~~vectors, with m the number of features/parameters of the data, measurements not affected by poor back-scattering or noise will cluster together in regions of high data density, as shown in Figures 4 and 3. The approach presented here identifies such clusters applying DBSCAN on data described by CNR, V_{LOS} and, additionally, spatial location and smoothness features, which help to make clusters more distinguishable.

DBSCAN identifies clusters and noise based on two parameters~~, the~~: a neighbourhood size, ε, and ~~the~~ a minimum number of nearest neighbours, ~~NN~~NN. The parameter ε is the euclidean distance from one observation to the limits of a neighborhood ~~in which NN~~ that might contain NN (or more) nearest neighbors~~of each point may be located~~. Intuitively, these parameters will define the minimum density that a ~~data partition~~ group of data points needs to have to be identified as a cluster. ~~In formal terms, clusters are defined as a collection of *density connected points*, the most general category out of four:~~

Observations within a cluster fall into the following categories,

- Core point: ~~A point~~ points q ~~is a core point if within its~~ whose ε-neighborhood ~~we can find NN~~ contains NN or more points~~apart from q~~.

- Direct density reachable point: ~~A point~~ points p ~~is directly density reachable from a core point~~ which are reachable by q ~~if p is in the~~ by laying within its ε-neighborhood~~of q~~.

- Density reachable point: ~~A point~~ points p ~~is density reachable from~~ reachable by a point r ~~if there exist~~ through one or a set of ~~core points q directly connected between them and to r and p.~~

- ~~Density connected points: Points p and r are density connected if they are density reachable for at least one common core point~~ directly connected core points q.

~~The euclidean distance here is m-dimensional, because the data is characterized by m features. Since the values taken by each feature can be very different (V_{LOS} is in a range from -35 to 35 m/s, and range gates can be between 105 and 7000 m apart from the lidar, for instance), each feature of the data needs to be centered and scaled properly before the application of~~ DBSCAN ~~to obtain a meaningful distance between points. By using the mean value for centering and the standard deviation for scaling, this step is very sensitive to the presence of outliers. Therefore, in our case, the mean is replaced by the median and the scaling is done using the inter-quartile range instead, which is the range between the 25th and 75th percentiles.~~

~~(a) DBSCAN algorithm definitions: direct density reachable point p (reachable by the core point q) and density reachable and density connected points p and r. Here point n does not belong to any of these categories, but noise. The DBSCAN algorithm working: (b) The current point being evaluated have the minimum number of nearest neighbours required, NN, within a neighborhood of size ε, classified as a *core point* (red) (c) The next point have less than NN neighbours, but one of them is a core point and becomes a *border point* (yellow) (d) A point with neither NN neighbours, nor core points within ε, classified as *noise* (brown) (e) The final cluster and noise. The former is a collection of density connected points.~~

~~Figure 5 shows schematically how this algorithm works. After centering and scaling, all m-dimensional distances between points are calculated, and DBSCAN starts traveling across all~~ travels across data points identifying ~~them with the categories already presented:~~ core points, ~~or points with NN neighbours within the ε-neighbourhood, border points , or~~ border points (density reachable points ~~that do not meet the NN requirement but still have~~ with at least one core point within the ε-neighbourhood~~, and points classified as~~) and noise, or points that ~~are not density connected~~do not belong to any of the categories described above. Finally, the algorithm ~~define clusters as the~~ separates clusters as individual groups of ~~data with only~~ density connected points~~and an unique group with points classified as noise, not belonging to any cluster. The~~ . Figure 5 shows schematically these definitions and how the algorithm works.

The input parameters ε and ~~NN~~ NN have a significant influence on the number and characteristics of the clusters detected~~by DBSCAN. Large values of~~ . For example, large ε together with ~~small NN will define very sparse clusters , including noiseas valid cluster members. On the contrary, the requirement of small, densely populated neighborhoods will reject many points that otherwise are valid cluster members. As we can see, both parameters are closely related, and therefore one can fix one of them to vary the one remaining. Following the guidelines of the author of DBSCAN in Ester et al. (1996) NN is fixed to 5 neighbours in this case, no matter the type of data or its distribution, leaving only ε to be estimated according to the structure of the data~~ a small NN will end up with sparse clusters that might include noise. ~~One way to describe this data structure is via~~ In order to find the parameters separating the least dense cluster from noise, we fix NN to a certain value k and determine ε from the data density distribution. The latter is well described by the *k-distance* function, $d_k(n)$~~. This function maps and sorts in ascending order the distance from each data point to its~~ , which represents the distances from all data point n to their

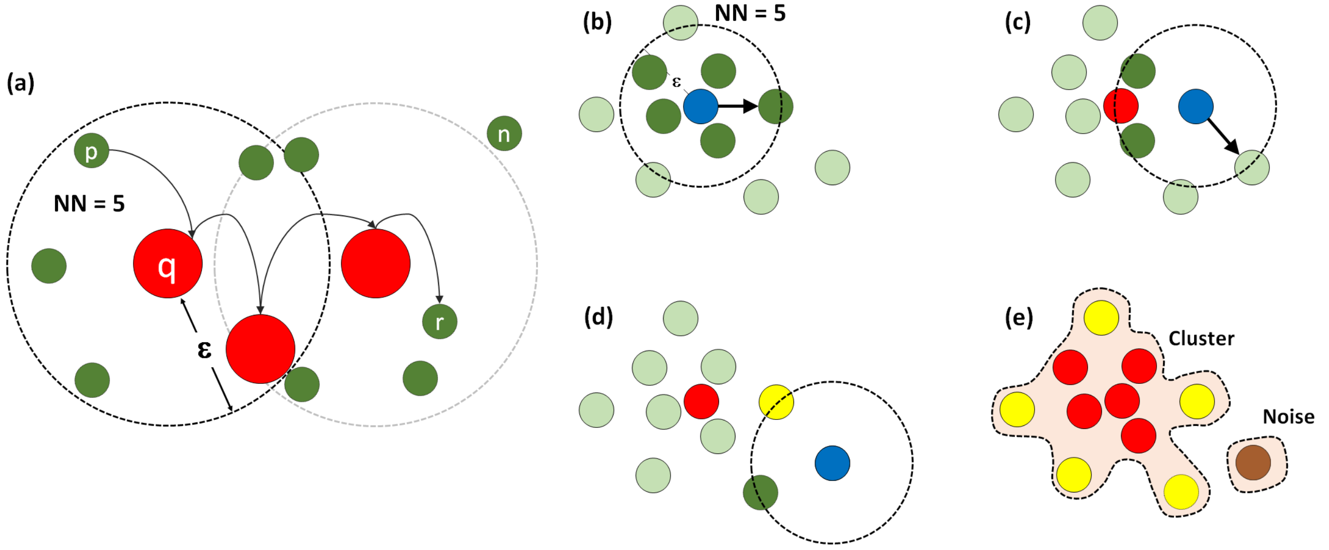

Figure 5. (a) DBSCAN algorithm definitions: direct density reachable point *p* (reachable by the core point *q*) and density reachable and density connected points *p* and *r*. Here point *n* does not belong to any of these categories, but noise. The DBSCAN algorithm working: (b) The current point being evaluated have the minimum number of nearest neighbours required, NN, within a neighborhood of size ε, classified as a *core point* (red) (c) The next point have less than NN neighbours, but one of them is a core point and becomes a *border point* (yellow) (d) A point with neither NN neighbours, nor core points within ε, classified as *noise* (brown) (e) The final cluster and noise. The former is a collection of density connected points.

respective k-th nearest neighbour~~. Figure ?? (a) gives an idea of how this function looks when applied on real data. When the distance to the fifth nearest neighbour (~~, sorted in ascending order. When k ~~= NN =~~ is 5 ~~) is considered we can distinguish four turning points, or **knees**: the first, positive, at very small distances, the second, highlighted , represents the limit~~ for instance, $d_5(n)$ in Figure 6 shows sudden changes (or knees) that give some indications about the data density distribution. The knee highlighted represents a limit between a group of reliable observations and ~~non-reliable, and two more, both in the region where k-th distance grows faster within the non structured group of data, identified as noise. These knees in *k-dist* represents sudden changes in the data density, and they separate clusters from each other and from mere noise. One way to determine the position of these knees is to locate the corresponding~~ the one growing fast towards noisy data. The positions of these knees in the graph correspond to the peaks in the curvature of the $d_k(n)$~~function~~, $\kappa(n)$ ~~. Even though $d_k(n)$ is discrete,~~ in expression (3)~~defines $\kappa(n)$ from its continuous analog, in which the~~ . In this expression primes correspond to the derivatives of $d_k(n)$ with respect to ~~the "continuous" data point number~~ a n. The continuous ~~representation of the discrete $k-dist(n)$ function~~ version of $d_k(n)$ is made by spline-fitting on a reduced set of uniformly distributed points over the original data set. ~~By doing this, one can avoid the selection of local spikes not representing the general structure of the data.~~

$$\kappa = (d_k(n))''/(1+(d_k(n))'^2)^{3/2} \tag{3}$$

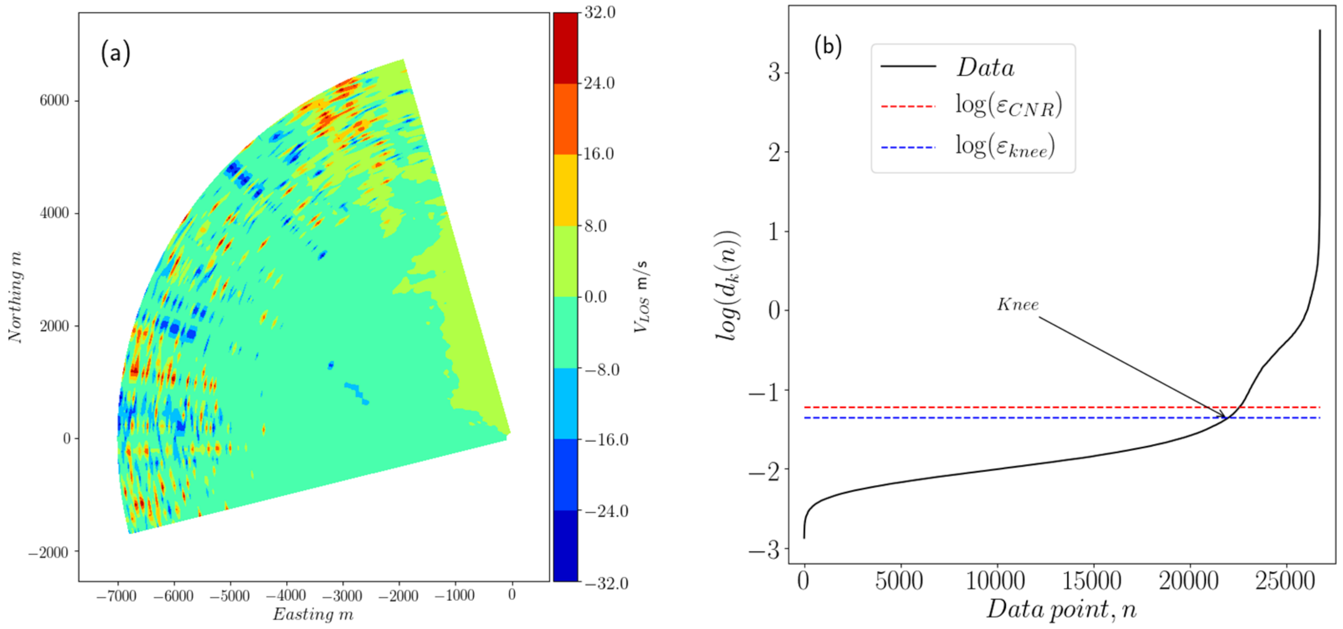

Figure 6. (a) Scan from the Balconies experiment (phase 1) with a 48% of data points in the range of reliable observations with CNR ∈ [-24, -8] dB (b) Logarithm of sorted distances to the 5-th nearest neighbour for each point in a data set. The ~~same scan after filtering~~ total number of observations corresponds to three consecutive scans, or 26730 points. The sorted 5-th distances show three knees separating three types of structures: reliable observations with ~~DBSCAN~~ ~~using V_{LOS}~~distances below ε_{knee}, ~~range gate~~an overlapping region where the distance between points grows faster, ~~azimuth angle, CNR~~ and ~~ΔV_{LOS} as features~~pure noise or non structured data.

~~For scans representing good measurements the relative distance between these knees is well marked and easy to identify. This is not the case when scans are very nosy, like the one in Figure 6, and the positions of knees become closer , with a large number of observations showing~~ When scans are very noisy, the selection of a proper value of ε is difficult, since knees are located closer together and a larger fraction of observations show a fast growing ~~k-th distance. The selection of a ε value that defines a reliable cluster of good observations is in this caseis difficult but can be eased when~~ $d_k(n)$, as expected. In this case, the fraction of points ~~with~~ showing a reliable CNR ~~value is also~~ values is taken into account ~~. In this case the neighborhood size is not selected by ε_{knee}, which is the ε-distance corresponding to the first noticeable knee from left to right, but by ε_{CNR}, defined by~~and ε is estimated by expression (4). Here f_{CNR} corresponds to the fraction of ~~very reliable observations in the data set (or measurements showing~~ observations CNR values within the range [-24, -8] ~~) over the total number of points. The~~ and the constants c_1 and c_2 are ~~defined by~~ obtained obtained from the upper and lower bounds of ε ~~, defined as the values that $k-dist(n)$ takes at the first and last knee from left to right in Figure 6 (a)~~in the data, respectively.

$$\varepsilon_{CNR} = c_1 f_{CNR} + c_2 \tag{4}$$

~~Data structure of the scan shown in Figure ??. (a) Logarithm of sorted distances to the 5-th nearest neighbour for each point in a data set. The total number of observations corresponds to three consecutive scans, or 26730 points. The sorted 5-th distances show three knees separating three types of structures: reliable observations with distances below ε_{knee}, an overlapping region where the distance between points grows faster and pure noise or non structured data. (b) Tri-dimensional representation of the data (range gate, CNR and V_{LOS}), where two coherent structures of data points, or clusters, are identified. The large structure has distances between observations below ε_{knee} and corresponds to reliable observations. The vertical line showing high CNR values are observations of bad V_{LOS} measurements in a group of 7 turbines around 2000 m from the lidar (see Figure 13).~~

~~(a) Scan from phase 1 of the experiment with a 13% of data points in the range of reliable observations with CNR $\in$ -24, -8dB (b) Data structure of the very noisy data. Here ε_{knee} over estimates the neighbourhood distance for a coherent cluster, and the inclusion of noisy measurements is avoided via ε_{CNR}~~

~~Regarding the features considered to characterize each data point, depending on whether we filter synthetic or real data , these will be radial or line-of-sight wind speed, V_{LOS},~~ The set of features considered when filtering synthetic data does not include CNR, because it is not available from the the lidar simulator described in Section 3. For synthetic and real data sets we consider spatial location (azimuth and radial positions~~and ΔV_{LOS}~~) and smoothness as additional features. The latter, ΔV_{LOS}, corresponds to the median difference in ~~radial wind speed of an specific radial and azimuth position with~~ V_{LOS} between a specific position and its direct neighbours ~~. This feature is included to consider the smooth spatial fluctuations of the radial wind speed, same assumption used by the median-like filter . The real data will include CNR as an additional feature. The controlled performance test on synthetic data does not include this feature, because the numerical lidar described in Section 3 just estimates V_{LOS}.~~ in one individual scan.

Since we consider features that vary importantly in magnitude (CNR and range gate distance for instance), we normalize the data before the application DBSCAN. This step is necessary for the estimation of meaningful distances between observations, basis of this approach. There are several ways to do this. Here, the data in each feature is centered by subtracting its median, and scaled according to its inter-quantile range. This aims to minimize the influence from outliers in the normalization.

The clustering filter is implemented to be a non-supervised classifier, and does not need more input parameters than the different features and the number of scans put together as a batch before filtering. The latter is set to three in this case, to speed up calculations and avoid creating clusters from noisy regions. From this point of view, this filter is also dynamic as that of Beck and Kühn (2017) when applied to a real data set, since it will consider the data structure within a period limited to 135 seconds (3 scans of 45 seconds in our case), and characteristics of temporal evolution of the data is indirectly taken into account. For the synthetic data used in this test, more than one scan filtered per iteration gives enough data density in noisy and reliable areas of the observational space. We speculate that scans that are correlated in time will enhance the self-similarity of the data, thus improving the performance of the filter. Turbulence structures with length scales in a range between the range gate size and the scanning area size will evolve at a slower rate than the time elapsed between consecutive scans.

5 Performance metrics

5.1 Synthetic data

~~The advantage of testing the filterson controlled cases is the~~ Expressions (5) to (7) defines three metrics to assess the performance of the filters, given prior knowledge of the position and magnitude of noise ~~. This allows us to define three metrics to assess how the median-like and clustering filters perform, namely, 1)~~ in a controlled case with N observations. The fraction of noise detected, η_{noise}, ~~2) fraction of good observations recovered η_{recov} and 3) a total performance metric η_{tot}, which takes into account the relative importance~~ quantify the relative importance of true positives, or the difference between observations identified as noise, N_{noise}, and false positives, N_{pos}, over the total number of contaminated observations~~, N_{cont}, and~~. The fraction of good observations recovered, η_{recov}, give an idea of the true negatives over the total number of non-contaminated observations, $N_{non-cont}$. True negatives are not not equal to $N - N_{noise}$, since the latter might include false negatives, N_{neg}. The relative importance of this two metrics, ~~using the noise fraction~~ for a given fraction of noise in a contaminated scan, f_{noise}~~as weight of the two metrics η_{noise}~~, is quantified by η_{tot}, which takes into account cases with a large fraction of noise detected and low recovery rate, and ~~η_{recov}, which are complementary as a large noise detection fraction will have associated a lower recovery fraction of good measurements. In expressions to this metrics are defined in formal terms, where N_{noise} is the number of observations identified as noise by the filter, N the total number of observations, and N_{pos} and N_{neg} the total number of false positives and false negatives, respectively~~vice-versa.

$$\eta_{noise} = \frac{N_{noise} - N_{pos}}{N_{cont}} \tag{5}$$

$$\eta_{recov} = \frac{N - (N_{noise} + N_{neg})}{N_{non-cont}} \tag{6}$$

$$\eta_{tot} = f_{noise}\eta_{noise} + (1 - f_{noise})\eta_{recov} \tag{7}$$

5.2 Real data

~~Unlike the controlled test, in real measurements we do not have any explicit reference to validate the results from the two filters. The~~ In the absence of reference measurements, the quality of the data retrieved after ~~their application~~ filtering is assessed by comparing the distribution of radial wind speeds for very reliable observations (with CNR values within the range between -24 to -8 dB) with the distribution of filtered observations that fall out of this range. Observations out of the reliable range population usually show a probability density function (or ~~pdf~~PDF) with heavier tails, like the pdfs in Figure 7. Here we understand a heavy tailed pdf as a distribution that slowly goes to zero and show higher probability density for values beyond the 3-σ limit (or 3 standard deviation limit), when compared to the normal distribution, evidence of a higher probability of occurrence of outliers or extreme values. The recovering rate of observations beyond the [0.003, 0.997] quantile range of the reliable V_{LOS} (shaded area in Figure 7) could shed information about the quality of the data retrieved by the filter.

Other metric is the similarity between pdf of reliable and non reliable data, after filtering. The distance between both probability density functions can be compared with similarity metrics like the Kolmogorov-Smirnov test (Kolmogorov, 1933) or Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). The former test measures the statistical similarity between two random variables, X_1 and X_2, by estimating the statistical distance, D (or K-S statistic), between their cumulative distribution functions, $F_1(x)$ and $F_2(x)$, as the supreme of their difference,

$$D_K = \sup_x \|F_1(x) - F_2(x)\| \tag{8}$$

The null hypothesis here is that two realizations are from the same distribution, if the K-S statistic is such that its two tailed p-value is above a certain level α. Due to the amount of data analyzed here is huge—we analyzed over 20000 scans for the two phases of the Østerild campaign, each with 8910 data points, over almost 10 days—this similarity test is very precise, but also very strict rejecting the null hypothesis for small deviations between $F_1(x)$ and $F_2(x)$. Nevertheless, the K-S statistics can be used to compare which probability distribution after filtering is closer to the one representing the reliable observations~~: the non-reliable observations after filtering with 1) the median filter or 2) the clustering filter approach~~.

The KL divergence is a measure of similarity, or overlapping of two distributions P_1 and P_2 , with realizations X_1 and X_2, respectively. It is used in different applications to shed light on the loss of information when X_1 is represented by P_2 or vice-versa and is defined by the expression (9).

$$D_{KL} = \sum_x P_1(x) \log\left(\frac{P_2(x)}{P_1(x)}\right) \tag{9}$$

Both metrics will be used to estimate how the distribution of non reliable observations of V_{LOS} is modified after filtering, and if the new distribution is similar (or close, in a statistical distance sense) to the probability density of reliable observations of the radial wind speed, shown in Figure 7 for phases 1 and 2 of the measurement campaign, respectively.

Both performance metrics, the recovery rate of abnormal measurements in the tails of the pdf of reliable observations and its statistical distance to the pdf of filtered non reliable observations, will be assessed for the median-like filter, the clustering filter and also for data filtered with a CNR threshold of -29 dB, following (Gryning and Floors, 2019).

6 Results

6.1 Synthetic data

~~As described in section 4.2, the median filter needs three parameters as input, n_r, n_ϕ and $\Delta V_{LOS,\,threshold}$, which will have a large impact on its outcome. For a given window size, both in radial and azimuth directions, a small value of $\Delta V_{LOS,\,threshold}$ will affect the capacity of the filter to retrieve observations that are valid but reflect large local wind speed fluctuations, affecting the data recovering rate. A larger value of this parameter in the other hand, will result in many noisy observations that are not the result of turbulent fluctuations being accepted as valid. The synthetic data set with 4305 scans and different turbulence characteristics allow us to test different combinations in these three parameters and find the optimal filter, which can be used~~

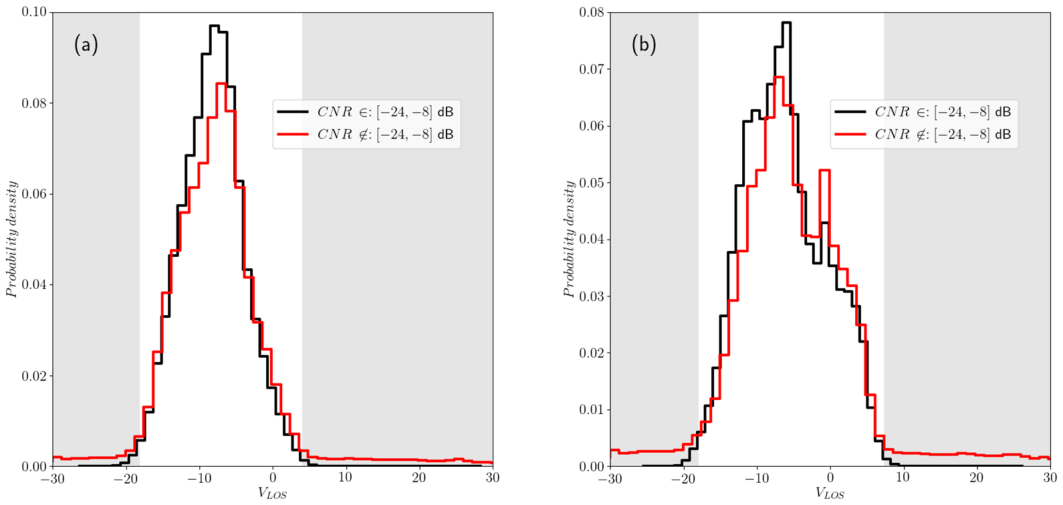

Figure 7. Probability density function of reliable observations of V_{LOS} (black solid line) and non reliable observations (red, solid line) for (a) Phase 1 of Balconies experiment with scans performed at 50 m a.g.l. and (b) Phase 2 of the same campaign, with scans performed at 200 m a.g.l.

~~to compare against the results of the clustering algorithm. The set of parameters used for this purpose are within the range of odd values from 3 to 13 elements for n_r, n_ϕ and in the interval 1 to 6 m/s for $\Delta V_{LOS,threshold}$. Figure A1 shows contours that present the most optimal value for η_{tot} among all possible values~~ In Figure 8 we can see the result of ~~$\Delta V_{LOS,threshold}$ and n_ϕ, for $n_r = 5$, the optimal window size in the radial direction. Large $\Delta V_{LOS,threshold}$ results in large η_{recov} but poor results for η_{noise} and the opposite for~~ the two filters applied on one synthetic scan contaminated with procedural noise. The contaminated observations are indicated by the grey area in this scan. Extreme values contaminating V_{LOS} are identified by both filters without problems, but subtle alterations on the original values of the ~~threshold, as expected. The metric η_{tot} then becomes relevant then to determine the optimal combination of parameters. From the contours it is possible to see that the performance in terms of the η_{tot} metric is less sensitive to n_ϕ than $\Delta V_{LOS,threshold}$. Even though the results here show average metrics for all the scans simulated, the optimal value of $\Delta V_{LOS,threshold}$ increases with the turbulence energy and length scale parameters, which is problematic, because it requires previous knowledge of turbulence characteristics that usually are not available before reconstruction, and more important, data filtering. In order to compare the performance of the~~ scan are hard to detect for the median-like filter~~to the clustering filter, the optimal set $n_r = 5$, $n_\phi = 3$ and $\Delta V_{LOS,threshold} = 2.33$ m/s will be used.~~

~~Contours of performance metrics for $n_r = 5$ over the $\Delta V_{LOS,threshold}$-n_ϕ space. Each point in in the contour plot corresponds to the mean value of (a) η_{noise}, (b) η_{rec} and (c) η_{tot} among all the 4305 synthetic scans filtered. The optimal value corresponds to $n_r = 5$, $n_\phi = 3$ and $\Delta V_{LOS,threshold} = 2.33$ m/s~~

. The clustering filter ~~uses V_{LOS}, the azimuth and radial positions, r and ϕ, and ΔV_{LOS}. Due to the nature and simplicity of the numerical lidar implemented, it is not possible to generate synthetic values of CNR, and this feature can not be included to characterize the data points. The clustering filter is implemented to be a non-supervised classifier, and does not need more~~

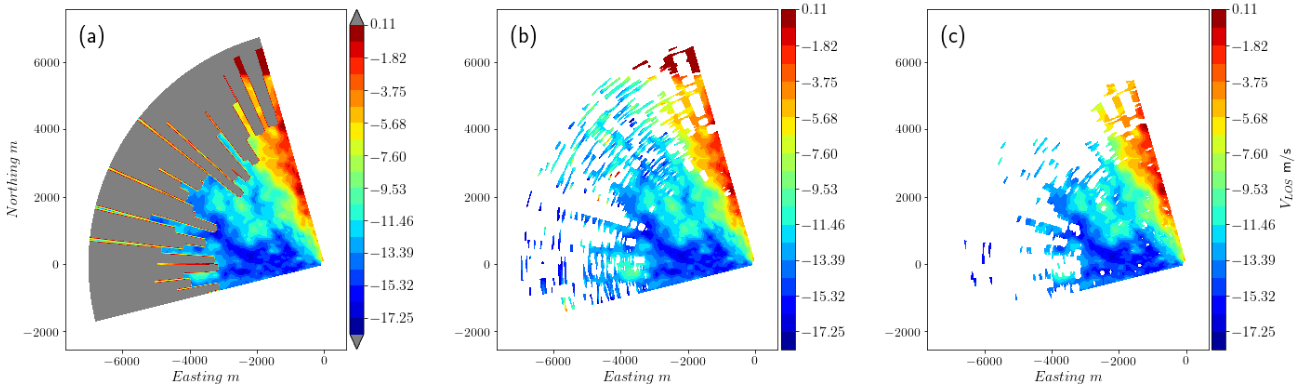

Figure 8. (a) Contaminated synthetic scan with noise indicated by grey area. (b) Scan filtered using the median-like approach. (c) Clustering filter.

~~input parameters than the different features and the number of scans put together as a batch before filtering. The latter is set to three in this case, to speed up calculations and avoid creating clusters from noisy regions. From this point of view, this filter is also dynamic as that of Beck and Kühn (2017) when applied to a real data set, since it will consider the data structure within a period limited to 135 seconds (3 scans of 45 seconds in our case), and characteristics of temporal evolution of the data is indirectly taken into acceoount. For the synthetic data used in this test, it is useful to have more than one scan filtered per iteration and have enough density of data points in both, noisy and reliable areas of the observational space. We speculate that scans that are correlated in time will enhance the self-similarity of the data, thus improving the performance of the filter. Turbulence structures with length scales in a range between the range gate size and the scanning area size will evolve at a slower rate than the time elapsed between consecutive scans. These structures will be then present at positions that are not very distant from its previous location in the earlier scan. The consequence of this is more dense clusters for good observations than when we consider independent scans realizations.~~

performs very efficiently detecting this type of contaminated observations and filters almost all the noise. Both filters repeat this behavior in all the synthetic scans used for this controlled test, as can be seen in Figure 9, which shows the resulting metrics of the two filters applied on the whole synthetic data set. Looking at η_{tot}, both filters show similar mean values and spread, with the clustering filter performing slightly better. The difference becomes noticeable when we see η_{noise}, which for the clustering filter show a mean value of 0.95, far larger than the 0.67 of the median-like filter. The latter result could be problematic if the median-like filter is used, since noise contaminating the filtered scan will result in non realistic wind fields after reconstruction.

Both filters perform well when evaluated in terms of η_{rec}, with the median-like filter showing a higher mean fraction of good observations retrieved, 0.96, compared with the 0.89 of the clustering filter. This result is expected, since the median-like filter is more permissive regarding fluctuations that can seem locally anomalous for the clustering filter. ~~It is not clear that the recovering rate of the clustering filter will be benefited by including more features from the data set. The euclidean distance used by the clustering algorithm to identify nearest neighbours increases (to a certain level) as we add more features or~~

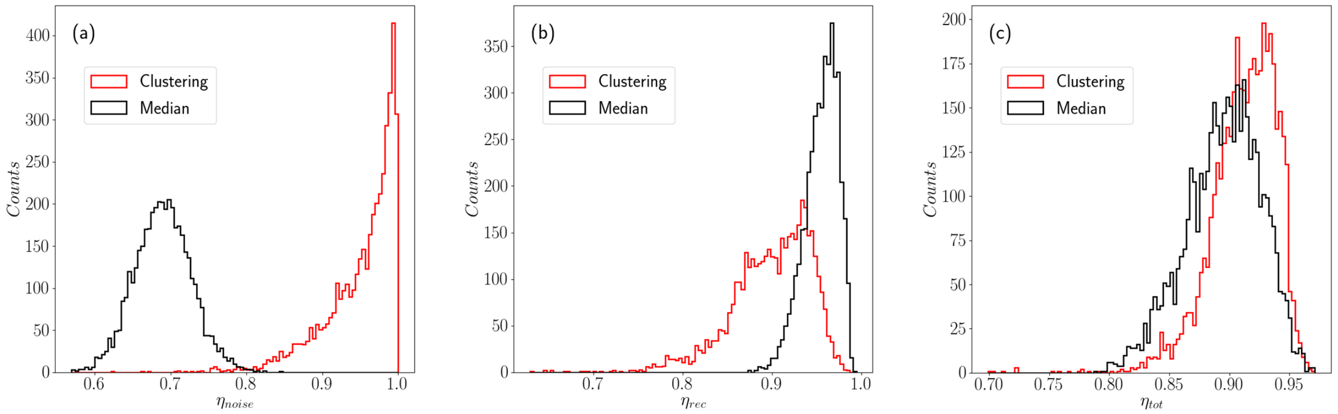

Figure 9. Histograms of the three performance indexes for the total number of synthetic scans (a) Both filters show similar spread but the clustering filter rejects a rather higher fraction of noise. (b) The higher recovery rate of the median-like filter, and its narrower distribution is superior than the clustering algorithm, the cost is acceptance of more contaminated observations (c) Both filters have similar mean values for η_{tot} around 0.9

~~dimensions the data description. As a consequence, adding more information to the data set, like CNR for instance, will result in less dense clusters and noise, making it easier to identify noise and improving the noise detection ratio, but on the other hand making the identification of good measurements located near the border of a cluster more difficult, especially if the CNR value of a good measurements is too low/high, due to poor scattering/closeness to hard targets. This might be solved including more scans per filtering iteration, but this can not be tested with the synthetic data, given the simplified numerical lidar implemented here.~~

6.2 Real data

The data set from the Balconies experiment presents advantages for the clustering filter, since the CNR value can be included as a feature in describing the data. Nevertheless, as mentioned already in section 2, we do not count on any reference to asses the performance of the filter apart from the radial wind speeds distribution of very reliable observations with CNR values within the range between -24 dB and -8 dB. As mentioned earlier, valid observations in this range might present a similar distribution. Figure 7 shows this distribution before filtering, shadowing the area of values of V_{LOS} that fall in the region beyond a 99.7% of the total probability or 3σ limit, usually classified as outliers. Figures 10 and 11 show the recovery fraction for CNR, median-like and clustering filters when applied on data in the reliable and non reliable CNR ranges for phases 1 and 2 of the Østerild experiment. Unlike the clustering filter, the CNR threshold and median-like filters show non negligible recovery rates beyond the 3-σ limit, particularly significant in the former. This result is very much in line with the η_{noise} metric from the synthetic data. Within the 3-σ range, the CNR and median-like filters perform slightly better than the clustering filter in terms of recovery fraction, in agreement with the results of η_{rec} in section 6.1. Even though this might compensate the fact that CNR threshold and median-like filters fail to filter out the major part of outliers, increasing the availability of measurements, this difference

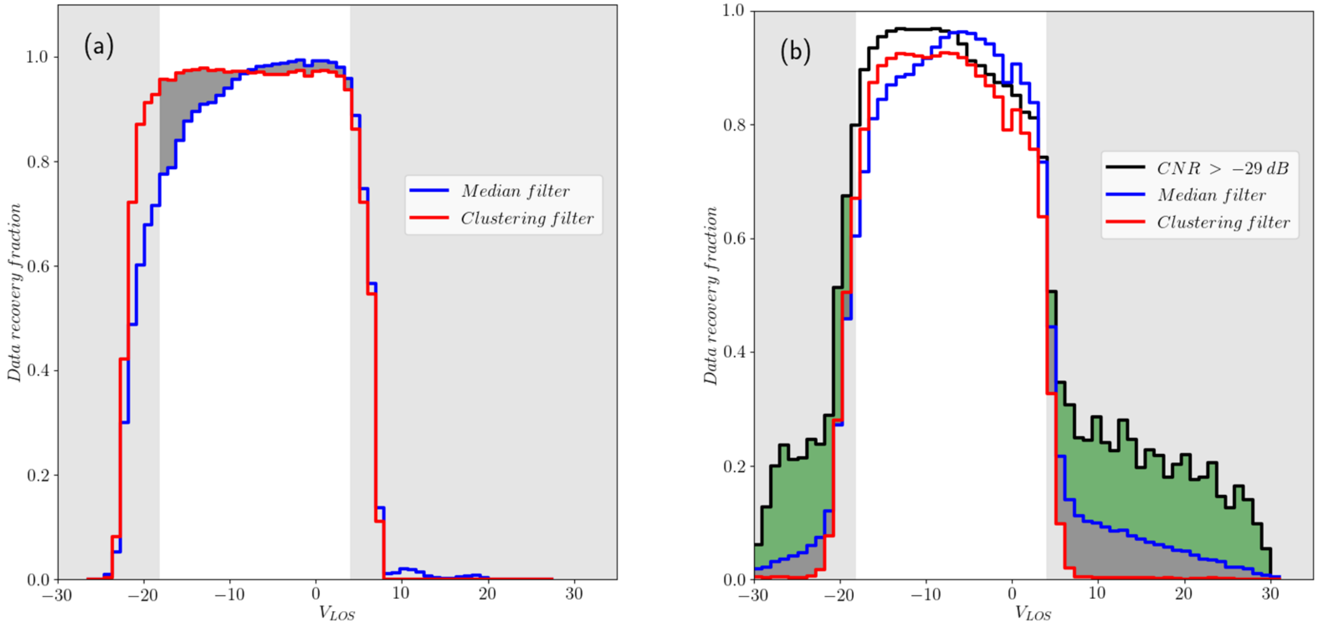

Figure 10. Distribution of recovery fraction per wind speed bin for phase 1 of the experiment of (a) reliable observations ($-24 <$ CNR < -8) and (b) non reliable data (CNR < -24 or CNR > -8) for the three types of filter. The shadowed area in both graphs corresponds to the region where observations exceed the 99.7% of probability (or 3-σ limit) in the pdf of reliable observations. The darker shadowed areas highlights the additional fraction of extreme values non-filtered by the median-like and CNR filters, when the former uses the optimal input set n_r = 5, n_ϕ = 3 and $\Delta V_{LOS,\,threshold}$) = 2.33 m/s.

does not make the pdf of the filtered data more similar to the pdf of reliable data, as Table 4 shows via the metrics D_K and D_{KL}. According to this metric, the pdf of the data after the application of the clustering approach looks more statistically similar to reliable observations. This table also show D_K and D_{KL} of the non reliable data before filtering, which in all cases is improved, except for D_K for median and CNR threshold filters during phase 2.

550 Figures 12, 13 and 14 show the performance of the three different filters in different regions of the scan, from respectively phase 1 and 2 of the experiment. When the spatial distribution of the recovery fraction is analyzed, we can see that the lowest values shown by the clustering filter are mostly located in the far region of the scan which, in general, presents low CNR values. The spatial recovery rate during phase 1 also show that the median-like and clustering filters are able to identify hard targets, which are also a source of bad observations. For scans recorded at 50 m above ground level in phase 1, back-scatter
555 is affected by a group of seven turbines located approximately in the middle of the scanning area, with one turbine touching the end of the southern beams of the scan and a meteorological mast located very close to the lidar. Figure 13 shows a detail of the recovery rate associated with the flow in the vicinity of the turbines group, in which we can see that the clustering filter is able to identify better the turbine locations, recovering more data in the ~~sorroundings~~ surroundings when compared to

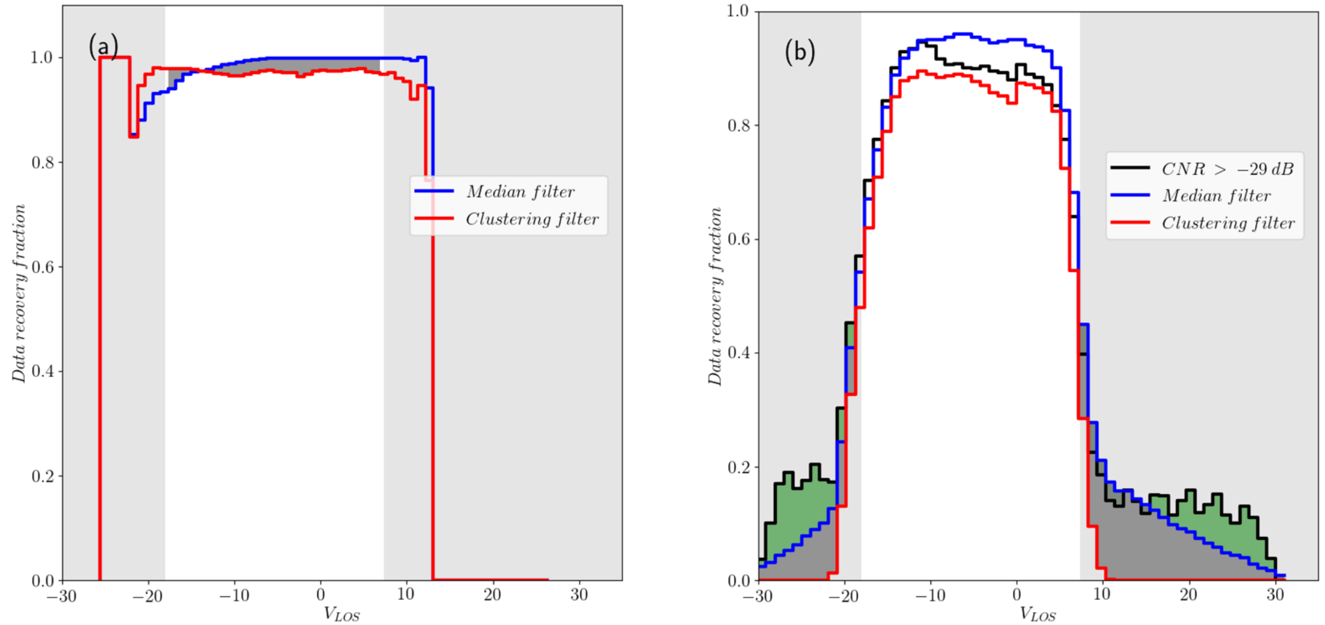

Figure 11. Distribution of recovery fraction per wind speed bin for phase 1 of the experiment of (a) reliable observations and (b) non reliable data for CNR, median and clustering filter. The shadowed area in both graphs corresponds to the region where observations exceed the 3-σ limit in the pdf of reliable observations. Again, darker shadowed areas highlights the additional fraction of extreme values non-filtered by the median-like and CNR filters, when the former uses the optimal input set n_r = 5, n_ϕ = 3 and $\Delta V_{LOS,\,threshold})$ = 2.33 m/s.

the median-like filter. The PDF of V_{LOS} in this area also show more similarities between the data filtered with the clustering algorithm and observations with CNR values in [-24,8].

Table 5 shows a summary of the additional data available when the CNR = -29 dB threshold, the median-like and the clustering filters are applied instead of the more conservative and restrictive CNR = -24 dB threshold filter. Additionally, this table shows the fraction of observations exceeding the 3-σ limit that are recovered by the three filters. Even though the clustering filter shows a slightly lower fraction of additional data available when compared to the other filters, most of it comes from values within the 3-σ region. Moreover the quality of the data recovered by the clustering approach seems to be higher when all these results are tested with the performance metrics defined in Section 5.

7 Discussion

7.1 ~~Performance assessment on synthetic data~~

The metrics introduced in section ~~6.1~~ 5.1 attempt to evaluate two different capabilities of the filters: the quality ~~of the data recovered and the amount of good quality~~ and amount od the data recovered. In general these two metrics are in

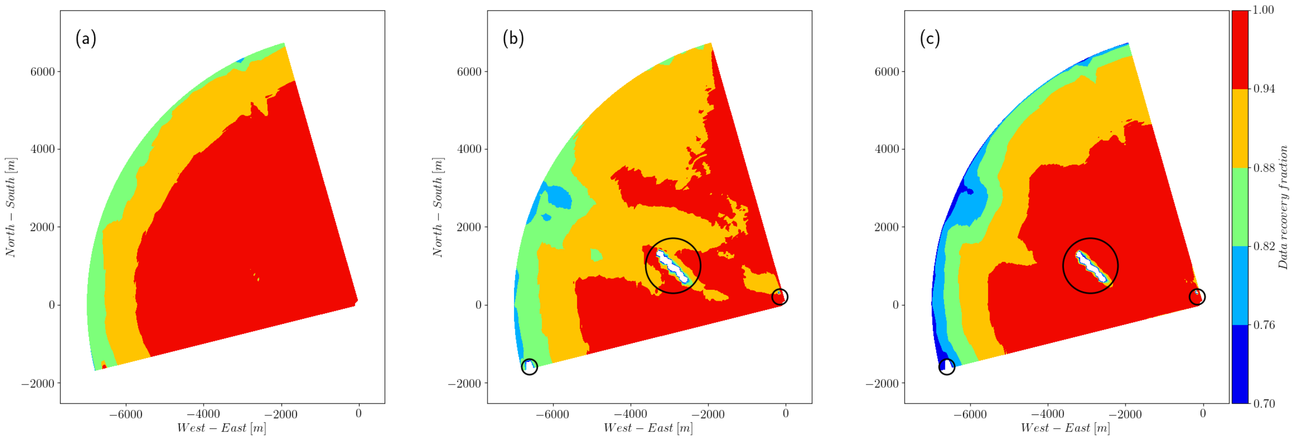

Figure 12. Total recovery fraction for phase 1 of the experiment. The noisy and far region of the scans show a high recovery, above 80%, for (a) the CNR > -29 dB threshold filter and (b) the median-like filter and below 75% for (c) the clustering filter. Highlighted, it is possible to see three groups of hard targets (turbines and one meteorological mast, close to the lidar), which are identified by the median and clustering filter with recovery rates below 20%.

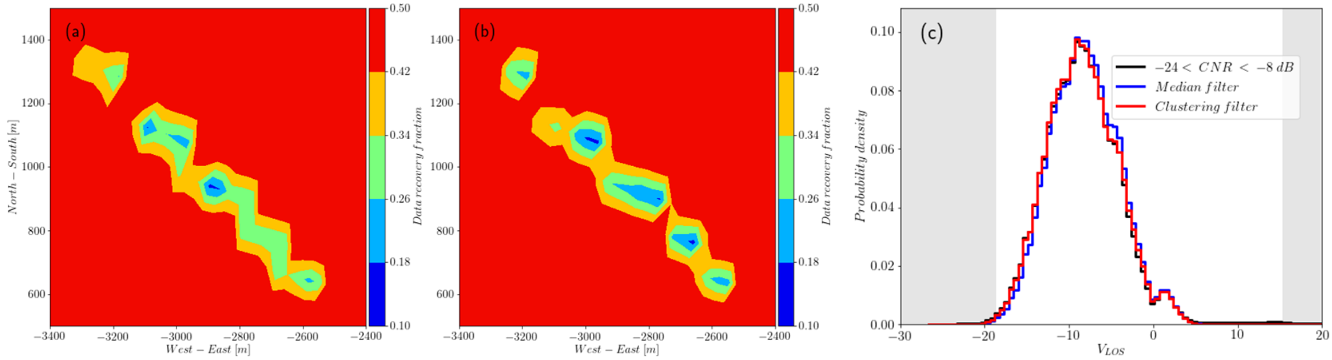

Figure 13. Detail of the recovery rate at the site of the turbines for (a) median filter and (b) clustering filter. The recovery is lower in the flow regime of the turbines cluster (there are 7 turbines in line) and higher in their surrounding for the clustering filter. Red denotes recovery rates of 0.5 or higher.(c) Probability density of V_{LOS} around the group of turbines

~~conflict—specially for the median-like filter—every~~ conflict, every time a high rate of noise ~~is removed also good measurements will be removed~~will decrease the data recovery. The metric η_{tot} attempts to quantify their relative importance regarding the noise fraction, which in this study is distributed in a relatively wide range, but on average represents 20% of the total number of measurements per scan. The impact of the noise fraction distribution on the performance of the filters was not explored, and variations on its dispersion and mean value might be necessary. Regarding the synthetic scans, they do not allow the identifi-

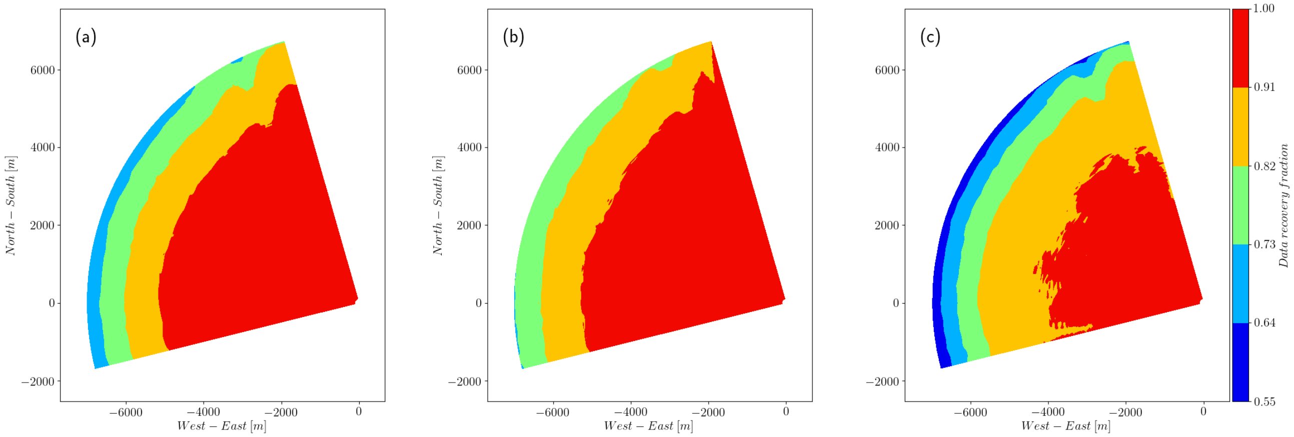

Figure 14. The total recovery fraction of observations for phase 2 of the experiment. The noisy and far region of the scans show a high recovery, above 70% for (a) and (b), the CNR > -29 dB threshold and median-like filters, respectively. The recovery decreases to 55% in the same region for the clustering filter, in line with the previous results, assuming that outliers (above the 3σ limit) and noise are more likely to be located here.

Table 4. Results of pdf similarity test of reliable and non-reliable data after filtering. The CNR = -29 dB threshold is also included~~by~~ (Gryning and Floors (2019))

Phase 1	D_K	D_{KL}
Non-reliable data before filtering	0.097	0.134
CNR threshold > -29 dB	0.045	0.109
Median filter	0.047	0.126
Clustering filter	0.037	0.105
Phase 2		
Non-reliable data before filtering	0.110	0.126
CNR threshold > -29 dB	0.114	0.052
Median filter	0.117	0.057
Clustering filter	0.103	0.045

cation of outliers in the time domain because they are time independent. A time evolving synthetic turbulence fields would be necessary to generate scans correlated in time and enhance the self similarity of the data. This might improve the performance of the clustering approach and allow the addition of a time dependence in the median-like filter, used already in Meyer Forsting and Troldborg (2016).

Table 5. Additional data recovered, relative to the amount of observations in the reliable range of CNR, and fraction of data recovered with values beyond the 3-σ range.

Phase 1, 3σ quantiles = [-18.16, 3.96] m/s	Fraction of data recovered beyond 3-σ	Additional data recovered
CNR threshold > -29 dB	27.1%	23.4%
Median filter	14.0%	23.1%
Clustering filter	8.6 %	22.1%
Phase 2, 3σ quantiles = [-18.08, 7.35] m/s		
CNR threshold > -29 dB	16.5%	40.4%
Median filter	12.6%	42.4%
Clustering filter	3.2%	38.1%

The synthetic wind fields used here do not consider the presence of hard targets. These anomalies in the wind field are observed by lidars as points with high CNR values and abnormal V_{LOS}. Assessing the performance of the filters in detecting such anomalies needs a more realistic model of the pulsed lidar. This ~~numerical lidar~~ lidar simulator would allow the generation of information normally available in real lidar measurements, like CNR, and the spread in the power spectra of the heterodyne signal, S$_b$. This additional information will benefit the performance assessment of the clustering filter and the simulation of hard targets. A more realistic lidar model was already implemented by Brousmiche et al. (2007), which can be used to explore further these aspects of the filtering process.

7.1 ~~Performance assessment on real data~~

The data set analyzed from the Balconies experiment corresponds to horizontal scans at 50 and 200 m above the ground level, limiting the analysis to one scanning pattern. Different scanning patterns, in vertical and horizontal planes, as well as wind fields over different topography would make this analysis more general, thus shedding light on the capabilities of the filters here presented. This is specially critical regarding the median-like filter, which might require again a sensitivity analysis to select proper parameters that adapt to different scanning patterns and turbulence field characteristics. So far, $\Delta V_{LOS,\, threshold}$ showed a dependence on the L and $\alpha\varepsilon^{2/3}$ parameters during the sensitivity analysis presented in Section 6.1. Larger fluctuations in the V_{LOS} field, whether they come from larger turbulent structures or higher turbulence energy or both, will need a larger value of $\Delta V_{LOS,\, threshold}$ to avoid the rejection of good measurements. Range Height Indicator (RHI) scanning patterns can pose the challenge of strong vertical shear and small turbulent structures that will need to reduce the window size n_r and n_ϕ for the median-like filter, and the selection of a different set of features (or a new definition for ΔV_{LOS}) for the clustering filter, in order to keep reliable observations from being filtered out.

Regarding feature selection, the clustering filter could consider the spectral spreading of the heterodyne signal, S$_b$ and time variation of V_{LOS}, in addition to features already used in this work to characterize and distinguish better cluster of good measurements. Nevertheless, due to the Euclidean distance definition, additional dimensions will make the data more sparse in higher dimensions, making it necessary to use more data points per filtering step (here we used only 3 scans at a time) to avoid the identification of good observations as spread, low density noise. It is because of this that the application of a feature selection method might be necessary (Chandrashekar and Sahin, 2014).

~~Finally, using~~ Using the statistical distances D_K and D_{KL} as a metric for the filter performance might not be totally correct. At range gates far from the lidar, the distance between beams increases, as well as the area covered by the accumulation of spectral information in azimuth direction. Averaging V_{LOS} over larger areas as we move forward through each beam, might affect the statistics and the ~~pdf~~ PDF of V_{LOS} (specially its spread) in the outer region of the scan. The fact that this region is where we usually find the non reliable measurements group~~(the one having CNR values out of the range between -24 and -8 dB)~~, may make the ~~pdfs~~ PDFs of reliable and non reliable observations somewhat different. ~~This~~ These possible deviations need to be investigated further.

7.1 ~~Advantages and limitations of proposed filters~~

~~From the results obtained in the analysis of real and synthetic data, the clustering filter show in general a better performance in noise removal and the recovering of good quality data from regions in the scan with poor CNR values. Moreover, this filter is based in a non-supervised clustering algorithm and requires little intervention from the user to obtain reliable results. This is a step forward to a more robust and automated processing of data from lidars, which ideally should be independent of the turbulence characteristics of the measured wind field or the scanning pattern used. The latter should be tested on a different data set, as mentioned earlier.~~

The selection of features and the amount of scans put together per filtering step/iteration could also be automatized, using feature selection methods. Nevertheless, this would make the clustering filter more complex in its implementation and more computationally expensive, which is the main disadvantage of this approach compared to the median-like filter. Very efficient median filters can achieve a computational complexity up to $\mathcal{O}(n)$, with n being the number of observations in the data set. Depending on the data structure, DBSCAN shows a computational complexity from $\mathcal{O}(n\log(n))$ to $\mathcal{O}(n^2)$. If the distance between points is in general smaller than ε, the first limit can be achieved, but clusters with different densities makes the algorithm less efficient. In the data analyzed here, having clusters with different densities is not an issue. Nevertheless, for non homogeneous flows, scans might persistently show regions with V_{LOS}, CNR or other feature with noticeable different values, may need to revisit the clustering algorithm used and implement a ε-independent clustering approach, like OPTICS (Ankerst et al., 1999) for instance.

8 Conclusions

The CNR threshold filtering has been the common approach to retrieve reliable observations form lidars measurements. In this work we compared this approach against two alternative techniques: a median-like filter, based on the assumption of smoothness of the wind field, hence, in the smoothness of the radial wind speed observed by a wind lidar, and a clustering filter, based in the assumption of self-similarity of the observations captured by the wind lidar and the possibility of clustering them in groups of good data and noise. A controlled test was carried out on the last two approaches, using a simple ~~numerical lidar~~ lidar simulator that sampled scans from synthetic wind fields, later contaminated with procedural noise. The results indicate that the clustering filter is capable of detecting more added noise than the median-like filter, at a good recovery rate of non contaminated data. When the three filters are tested on real data, the clustering approach shows a better performance on identifying abnormal observations, increasing the data availability between 22% and 38% and reducing the recovery of abnormal measurements between 70 and 80% when compared to a CNR threshold. This is an important result, because increases the spatial coverage of the data which can be used later for wind field reconstruction and wind data analysis, specially in the far region of the scan, that covers the largest measured area.

Even though the median-like filter is computationally efficient, it needs an optimal definition of input parameters, which are dependent on the turbulence characteristics of the wind field. The clustering filter is more robust in this sense, because it is capable of automatically adapt its ~~input parameter~~ parameters to the structure of the data. This is a step forward to a more robust and automated processing of data from lidars, which ideally should be independent of the turbulence characteristics of the measured wind field or the scanning pattern used.

Code and data availability. The synthetic data and code is available at https://github.com/lalcayag/Lidar_filtering. Real data can be found in Simon and Vasiljevic (2018)

Author contributions. The author Leonardo Alcayaga implemented the analysis on synthetic and real data and wrote all sections of this manuscript

Competing interests. The author delcares not having any competing interest

Acknowledgements. I would like to thank Ioanna Karagali for comments and guide through the data from the Østerild Balconies experiments, Robert Menke for the first version of the median-like filter and Ebba Dellwik, Nikola Vasiljevic, Gunner Larsen, Mark Kelly and Jakob Mann for the very useful and clarifying feedback during the analysis and writing process.

Appendix A: Sensitivity analysis on median-like filter parameters

Figure A1 shows contours that present the most optimal value for η_{tot} among all possible values of $\Delta V_{LOS,threshold}$ and n_{ϕ}, for n_r = 5, the optimal window size in the radial direction. Large $\Delta V_{LOS,threshold}$ results in large η_{recov} but poor results for η_{noise} and the opposite for values of the threshold, as expected. The metric η_{tot} then becomes relevant to determine the

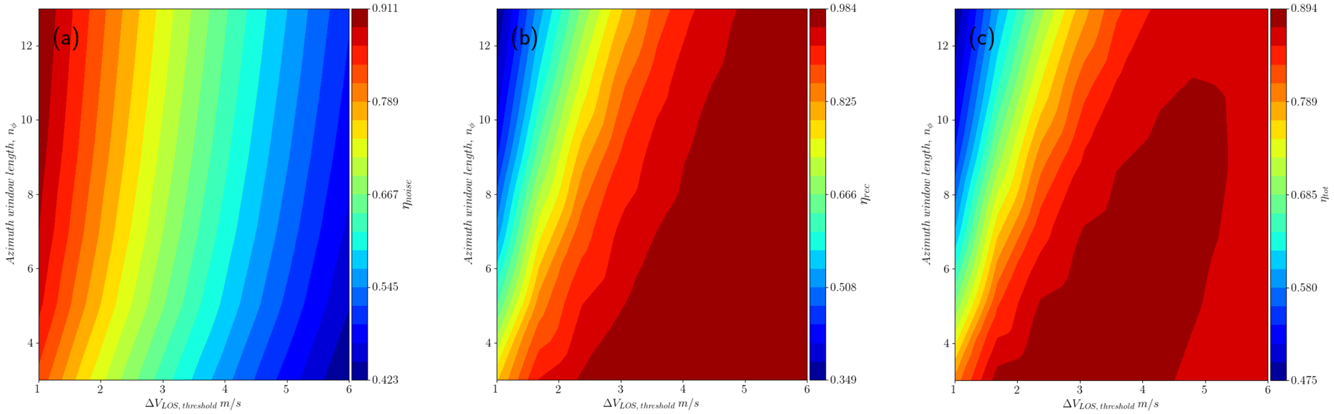

Figure A1. Contours of performance metrics for n_r = 5 over the $\Delta V_{LOS,threshold}$-n_ϕ space. Each point in in the contour plot corresponds to the mean value of (a) η_{noise}, (b) η_{rec} and (c) η_{tot} among all the 4305 synthetic scans filtered. The optimal value corresponds to n_r = 5, n_ϕ = 3 and $\Delta V_{LOS,threshold}$ = 2.33 m/s

optimal combination of parameters. From the contours it is possible to see that the performance in terms of the η_{tot} metric is less sensitive to n_ϕ than $\Delta V_{LOS,threshold}$. Even though the results here show average metrics for all the scans simulated, the optimal value of $\Delta V_{LOS,threshold}$ increases with the turbulence energy and length scale parameters, which is problematic, because it requires previous knowledge of turbulence characteristics that usually are not available before reconstruction, and more important, data filtering.

References

Ankerst, M., Breunig, M. M., peter Kriegel, H., and Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure, in: Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, pp. 49–60, ACM Press, 1999.

Backer, E.: Computer-assisted Reasoning in Cluster Analysis, Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK, 1995.

Banakh, V. A. and Smalikho, I. N.: "Estimation of the Turbulence Energy Dissipation Rate from the Pulsed Doppler Lidar Data, Atmos. Ocean. Opt., 10 (12), 957–965, 1997.

Beck, H. and Kühn, M.: Dynamic Data Filtering of Long-Range Doppler LiDAR Wind Speed Measurement., Remot Sens, 9(6), 561, https://doi.org/https://doi.org/10.3390/rs9060561, 2017.

Brousmiche, S., Bricteux, L., Sobieski, P., Macq, B., and Winckelmans, G.: Numerical simulation of a heterodyne Doppler LIDAR for wind measurement in a turbulent atmospheric boundary layer, in: 2007 IEEE International Geoscience and Remote Sensing Symposium, https://doi.org/10.1109/IGARSS.2007.4423420, 2007.

Burger, W. and Burge, M. J.: Digital Image Processing - An Algorithmic Introduction using Java., Texts in Computer Science, Springer, 2008.

Cariou, J.: Remote Sensing for Wind Energy, chap. Pulsed lidars, pp. 131–148, DTU Wind Energy, Denmark, 2015.

Chandrashekar, G. and Sahin, F.: A Survey on Feature Selection Methods, Computers and Electrical Engineering, 40, 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024, 2014.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, pp. 226–231, Portland, Oregon, 1996.

Gryning, S.-E. and Floors, R.: Carrier-to-Noise-Threshold Filtering on Off-Shore Wind Lidar Measurements, Sensors, 19, https://doi.org/10.3390/s19030592, 2019.

Gryning, S.-E., Floors, R., Peña, A., Batchvarova, E., and Brümmer, B.: Weibull Wind-Speed Distribution Parameters Derived from a Combination of Wind-Lidar and Tall-Mast Measurements Over Land, Coastal and Marine Sites., Bound-Lay Meteorol, 159(2), 329, https://doi.org/https://doi.org/10.1007/s10546-015-0113-x, 2016.

Huang, T., Yang, G., and Tang, G.: A fast two-dimensional median filtering algorithm, IEEE T. Acoust. Speech, 27, 13–18, https://doi.org/10.1109/TASSP.1979.1163188, 1979.

Karagali, I., Mann, J., Dellwik, E., and Vasiljević, N.: New European Wind Atlas: The Østerild balconies experiment, J Phys Conf Ser, 1037, 052 029, https://doi.org/10.1088/1742-6596/1037/5/052029, 2018.

Kolmogorov, A.: Sulla determinazione empirica di una lgge di distribuzione, Inst. Ital. Attuari, Giorn., 4, 83–91, https://ci.nii.ac.jp/naid/10010480527/en/, 1933.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, Ann. Math. Statist., 22, 79–86, 1951.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings Fifth Berkeley Symp. on Math. Statist. and Prob., vol. 1: Statistics, pp. 281–297, Berkeley, California, 1967.

Mandelbrot, B. B.: The fractal geometry of nature, W. H. Freeman and Comp., New York, 1983.

Mann, J.: The spatial structure of neutral atmospheric surface-layer turbulence, J Fluid Mech., 273, 141–168, https://doi.org/10.1017/S0022112094001886, 1994.

Mann, J.: Wind field simulation, Probabilist. Eng. Mech., 13, 269–282, https://doi.org/https://doi.org/10.1016/S0266-8920(97)00036-2, 1998.

Mann, J., Angelou, N., Arnqvist, J., Callies, D., Cantero, E., Arroyo, R. C., Courtney, M., Cuxart, J., Dellwik, E., Gottschall, J., et al.: Complex terrain experiments in the new european wind atlas, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 375, 20160 101, 2017.

Menke, R., Vasiljević, N., Wagner, J., Oncley, S. P., and Mann, J.: Multi-lidar wind resource mapping in complex terrain, Wind Energy Science Discussions, 2019, 1–21, https://doi.org/10.5194/wes-2019-85, https://wes.copernicus.org/preprints/wes-2019-85/, 2019.

Meyer Forsting, A. and Troldborg, N.: A finite difference approach to despiking in-stationary velocity data - tested on a triple-lidar, J Phys Conf Ser, 753, 072 017, https://doi.org/10.1088/1742-6596/753/7/072017, 2016.

Park, H.-S. and Jun, C.-H.: A Simple and Fast Algorithm for K-medoids Clustering, Expert Systems with Applications: An International Journal, 36, 3336–3341, 2009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Perlin, K.: Noise hardware, In Real-Time Shading SIGGRAPH Course Notes, 2001.

Rui Xu and Wunsch, D.: Survey of clustering algorithms, IEEE T. Neural Networ., 16, 645–678, https://doi.org/10.1109/TNN.2005.845141, 2005.

Simon, E. and Vasiljevic, N.: Østerild Balconies Experiment (Phase 2), https://doi.org/10.11583/DTU.7306802.v1, https://data.dtu.dk/articles/_sterild_Balconies_Experiment_Phase_2_/7306802, 2018.

Smalikho, I. N. and Banakh, V. A.: Accuracy of estimation of the turbulent energy dissipation rate from wind measurements with a conically scanning pulsed coherent Doppler lidar. Part I. Algorithm of data processing, Atmos. Ocean. Opt., 26, 404–410, https://doi.org/10.1134/S102485601305014X, 2013.

Stawiarski, C., Träumner, K., Knigge, C., and Calhoun, R.: Scopes and Challenges of Dual-Doppler Lidar Wind Measurements—An Error Analysis, Journal of Atmospheric and Oceanic Technology, 30, 2044–2062, https://doi.org/10.1175/JTECH-D-12-00244.1, https://doi.org/10.1175/JTECH-D-12-00244.1, 2013.

Vasiljevic, N., Lea, G., Courtney, M., Cariou, J.-P., Mann, J., and Mikkelsen, T.: Long-Range WindScanner System, Remote Sensing, 8, https://doi.org/10.3390/rs8110896, 2016.

Vasiljević, N., L. M. Palma, J. M., Angelou, N., Carlos Matos, J., Menke, R., Lea, G., Mann, J., Courtney, M., Frölen Ribeiro, L., and M. G. C. Gomes, V. M.: Perdigão 2015: methodology for atmospheric multi-Doppler lidar experiments, Atmospheric Measurement Techniques, 10, 3463–3483, https://doi.org/10.5194/amt-10-3463-2017, 2017.

Xu, D. and Tian, Y.: A Comprehensive Survey of Clustering Algorithms, Annals of Data Science, 2, 165–193, https://doi.org/10.1007/s40745-015-0040-1, 2015.