The reviewer's comments are given in green, while our response is provided in black.

The authors compare different methods for calculation of turbulence parameters from measurements with a single Doppler lidar. The topic is up-to-date and is very important for the further development of different scientific disciplines and the further technical development, e.g., of wind power plants. The authors have put considerable work into the paper and AMT is the right journal to publish this study. I recommend publication if the following major issues are addressed:

Major issues:
1.) At many points in the document, quite subjective descriptions of a correlation or a match are given. Please look into this issue. You could, e.g., quantify what you mean with "good", "bad", "show skill" or "accurate" once within the document and connect it with proper numbers. It will really upvalue the paper, if you make it more quantitative. It will help to transport the message.

In several places throughout the manuscript, we have added quantitative descriptions (i.e., correlation coefficients, slopes to indicate bias) to the text where it seems appropriate, including in the abstract. However, to ensure ease of the narrative when describing these data sources in the discussion/results sections, we found it preferable to use more qualitative language to indicate whether we thought these vales represented a good or bad fit with respect to the other techniques and measurements, and to avoid redundancy of the quantitative analysis already provided in the tables and figures themselves. Instead, the appropriate figures or tables are referenced in the text. In addition, Table 2 summarizes all the quantitative statistics for a quick look up by the reader. To further strengthen the quantitative analysis, we have also provided the RMSE values in the updated Table 2.

2.) p12 l10: -> This discussion must be mentioned earlier in the document. Is there really no way to estimate the uncertainties of these methods? I would doubt that. Understandably, analytic error calculation is extremely difficult for this kind of evaluations. However, numerical methods exist that can yield an error estimation for certain kinds of noise. You could, e.g, use Monte Carlo simulations, imposing small variations on the input data and then analyze in what range the results change. With such an approach it would be possible to discriminate between measurement errors and methodical uncertainties (e.g., incomplete overlap between the tower measurements and the lidar observations). That would greatly help the interpretation especially of Figs. 4/5 and 8/9. This topic is also connected with P.13 Line 8: "Approximately half of these outliers are negative TKE values, which were removed as discussed earlier..." -> Are those really outliers or just noisy values that happen to be close to zero. An uncertainty estimation or a more thorough description of the 6-beam technique would help here.

We agree with the reviewer that quantifying the uncertainty of the measurements is ideal for these intercomparisons and interpreting the results. Unfortunately, determining the magnitude of the uncertainty itself is not trivial. Entire studies have been devoted to addressing and quantifying sampling errors of turbulence and flux measurements from time-series analysis alone (see references in new discussion in paper). Since the measurement techniques presented here have are made using Doppler lidar observations over both time and space, no method has been developed yet to quantify errors with these techniques. Due to the intensive analysis required to properly attribute sampling errors from even one technique (as evidenced by multiple studies now referenced in the paper that evaluate sampling errors from time series analysis), entire studies could be done to determine the proper sampling errors for each of the techniques here separately. As such, properly quantifying and attributing sampling errors are out of the scope of this paper and is a topic for further studies. In lieu of quantifying these errors, we have added a more thorough discussion of sampling errors earlier in the paper before the results are presented (last paragraph before Sect. 4.1).

We do note that the input data (line-of-sight velocity measurements) are already contaminated with random error (i.e., noise). This is covered in Sect. 2 (see Eq. 1 & 2). However, using

established techniques using the autocovariance of the line-of-sight velocity measurements, we have been able to quantify and remove this noise from all of the scans except for the RHIs (as is discussed in Sect. 2). As such, random errors are not anticipated to be a significant source of error except for the RHIs, as discussed in the manuscript. With this approach, we feel that a Monte Carlo simulation is not appropriate to assign errors, as these noise effects are already quantified and removed.

The negative TKE values are simply outliers based on the definition used in the manuscript, being more than 1 order of magnitude difference between sonic anemometer and observations (see p. 13, line 20). As mentioned above, quantifying the sampling error of six-beam measurements is difficult due to the spatio-temporal nature of the technique and out of the scope of the paper. However, we do believe the negative values are thought to be related to the sampling error of the measurement (since there is no other plausible explanation), and state this where the negative values are first discussed on p.12 line 2.

3.) p9 l18: "These erroneous echoes were removed using a discontinuity-based algorithm described by Bonin and Brewer (2016)" -> Maybe it is not so easy. Such a correction is never perfect and some artifacts always remain. The kind of signal folding you experience imposes spatially confined biases on the measured signal (spanning some range gates). Some techniques may be more susceptible to this influence than others, introducing an unknown bias into the intercomparison. E.g., the six beam technique will be affected differently by spatially confined shifts in a single beam than the RHI scans. Signal folding is also no necessity for Doppler lidar measurements. They can be avoided by reducing the pulse repetition rate, which should be mentioned. Several other questions arise and have to be discussed: (a) How is it possible to identify the folding effect unambiguously in the data? (b) What percentage of data is affected? (c) What is the remaining bias after correction?

We agree with the reviewer that the cited technique to remove range folded echoes is not perfect, largely since it relies on contextual information. This technique works better on RHI and PPI scans than it does on measurements at single beam positions, due to the necessity of having data nearby spatially and temporally. While showing and discussing the results of the intercomparison, we specifically point out in the manuscript that the large number of outliers in the six-beam TKE measurements are largely a consequence of range folded echoes not being removed due to short amount of time at each beam position (see p. 13, lines 22-28).

We acknowledge that range folding does not appear in all Doppler lidar measurements. However, the commercial Doppler lidar systems that have been increasingly used by a diverse set of users in the past 5 years (Leopshere and Halo systems) have PRFs ~10-20 kHz and have been documented to be susceptible to range folded echoes (see Päschke et al. (2015) and Bonin and Brewer (2017)). By clearly stating (see p. 9, line 17) that the high PRF (20 kHz) of the system is why range folding is an issue, it is implied that range folding is not an issue for low PRF systems. Thus, we do not see a need to expand upon the issue in the manuscript.

While these anomalous echoes do affect the data quality and we feel that they are important to mention, they are not the main focus of the manuscript and much of the data (as determined by manual inspection) is not affected after the described and referenced QC methods are applied. Going into details about these echoes, their characteristics, frequency of occurrence, etc. requires rigorous analysis itself and is outside of the scope of the paper. Instead, we have added a statement in the manuscript to refer the reader to Bonin and Brewer (2017) for these characteristics and how these anomalous echoes are identified.

We have provided a statement in the manuscript indicating that 5.6% of data points were flagged and removed, which we believe to be an overestimate (since we conservatively remove

suspect data points). To really address this issue, a low-PRF system, such as our HRDL, needs to be synchronized with the high-PRF system scanning the same positions at the same time, to accurately quantify the percentage of data from that high-PRF system that is range folded through a direct intercomparison of measurements along each beam. This was not done during XPIA, and to our knowledge has not been done any other time before. Thus, the exact percentage of data affected by range folding is not known, until we can perform such as experiment.

Minor issues:
p2 l7: "good": -> As described above, please quantify...
We have decided to simply remove the word ‘good’, and provide a reference for the reader if they are interested in the details of the results of the cited study.

p2 10 "the long time series of staring": -> Again, please quantify. It is actually a very good question what "long" means here. You correctly cite Lenschow et al. (1994) here, but leave the calculations to the reader. Please give a rough estimation of what "long" means in this context.
As we have rewritten much of the introduction, there is no obvious location to discuss this here. However, we have added a paragraph in the beginning of Sect. 4 discussing sampling errors, in which we give an estimation of how long a time series needs to be under convective conditions (1-2 hours) and stable conditions (5-10 min).

p1 l4: "trusted in situ instrumentation": -> I think I know what you mean, but please give a reference of what "trusted" means in this context. Do you mean something like "officially approved by a standardization institution"?
We have clarified in the introduction at p. 3 l 15 that sonic anemometers are a commonly used reference device, and have provided a reference to International Energy Agency (IEA) report that supports this claim. We have decided to keep the wording unchanged in the abstract (at p1 l4), as we feel it is not the appropriate location to give a reference for this.

p1 l12: "None of the methods evaluated were able to consistently accurately measure the shear velocity" -> Please discuss what accuracy is necessary to measure shear velocity "accurately". Which maximum error is allowed for which purpose?
As the reviewer states, the maximum error allowed for a measurement to be useful depends on the purpose or use of the measurement. We do provide a general discussion on how accurate measurements need to be depending on the application in the paragraph starting on p. 2 l. 33. But we do not believe that the essence of this study is to sort out whether a measurement is accurate enough for a given purpose. That is for users to decide based on their specific objectives. Instead, the objective here is to analyze the measurement techniques to determine systematic biases and if measurements are of any use at all to any audience. In the particular sentence referred to here, we state that stress velocity was not found to be accurate by any of the methods tried. This is true regardless of the use of the measurement as the correlation coefficient between the sonic and VAD or six-beam u-star was 0.171 and 0.147 respectively and the scatter was large across all values, indicating very little correlation, and that the lidar measured u-star was insufficiently accurate for any purpose.

p6 l7: Since data were collected at 2 Hz, two samples were collected 0.5 s apart -> Please decide between mentioning "2Hz" or "0.5s".
We have removed the 2 Hz statement and now are using only 0.5 s here.

p8 l4: SNR<-27 dB -> which definition of SNR is applied here?

A statement has been added here to clarify that the SNR values were taken as the carrier-to-noise ratio produced by the lidar manufacturer's processing algorithms. Since the processing algorithms are proprietary, we are unsure about the details of how the SNR/CNR is exactly calculated and the exact definition used. However, the values of SNR/CNR should be comparable for measurements from other Leosphere Doppler lidars.

p16 l1: "show skill" -> Please define "skill" together with the other descriptions.
We have added a definition of 'skill' here as showing correlation between the lidar and the sonic anemometer (reference instrument).

Typos:
p1 l11: Typo: "to biased" -> "to be biased"
Corrected
Table2: typo at "0.547nb"
Corrected