We thank the reviewer for the considerable time and effort put into reading and critiquing the manuscript. We believe that the revised paper has significantly improved by addressing the concerns of the reviewers. Below, we have included a response to each of the reviewer's comments, and our responses are shown in red.

Main comments:

(1)

In some cases in this investigation, the autocovariance technique results in a negative error. It is assumed that this is due to volume-averaging effects, i.e. caused by the limited spatial resolution of the lidar measurement, which can lead to too steep slopes in the turbulence spectra and, thus, to an underestimation of the variance. However, it is not clear to me if this assumption that a negative noise-error is related to this effect is physically based. Moreover, it is not clear from the shown spectra if the volume-averaging effect really has an impact on the analyzed measurements. In the results shown, it also does not clearly have a positive effect.

I would suggest either to define a lower threshold of 0 for the assumed noise error and not to discuss the volume-error in this way or to analyze possible occurring volume-errors more closely and to demonstrate the relationship between the latter and the 'negative-noise' error.

Within the updated manuscript, we have added a comparison of the sonic anemometer and Doppler lidar power spectral density for the examples shown within Sect. 3.2, by showing additional plots within Fig. 4 and a discussion of these plots. Also, taking into account another suggestion by this reviewer, we have used a five point moving window on the spectra. This makes the spectrum shown herein much more smooth, so it is easier to visualize and identify the effect of spatiotemporal averaging in the Doppler lidar spectra, such as in Fig. 8, b. After making these changes, we believe that it is clear that the 'negative-noise' error within the autocovariance fitting is due to the fact that the high frequency eddies are not properly resolved by the Doppler lidar, as indicated by the drop-off in spectral density at higher frequencies.

(2)

In the aims as well as in the conclusion, it is said that the "optimal parameters [i.e., lag time] that should be used when applying the autocovariance method" are determined.

I would have expected a comparison of variances against sonic-derived variances, as in Fig. 10, but using different lag times. Instead, the optimal lag time is determined by theoretical considerations mainly without demonstrating the validity of these considerations. Only in section 4.1, the ideal lag time determined in section 3 is shown for the sonic measurements, but only compared against a much larger lag time. What about a fixed lag times between 1 s and 10 s, e.g.? In their article, Lenschow et al. (2000) suggest to use "the first few lags" (p. 1333, last sentence in left column in their article).

I would suggest not to mix up the discussion of the temporal averaging of sonic measurements with the analysis of the lag times. Anyway, I don't understand why you analyze the noise error of the sonic measurements at all – the main conclusion from this is "that the sonic observations […] contain little noise" (clearly evident from the red line in Fig. 6). It would be better to do this for the lidar data.

While we did try to minimize the sampling differences between the Doppler lidars and the sonic anemometers, they are sampling volumes that are adjacent and not identical volumes (the center of the measurement gate for the OU DL is often several metres off from the measurement height for the sonic, which we have clarified in the updated manuscript). This will lead to some small discrepancies in the measured statistics. For this reason, we prefer to stick with using only the sonic data to analyze the number of lags to use, that way it is a direct comparison of variance estimates with no concern about the representativeness of the statistics or any differences in the statistics at slightly different heights. Also by using the sonic anemometer data, we are able to address a comment by reviewer 1 to systematically investigate the effect of the number of lags used on the accuracy of measurements at different heights.

We also used the sonic data to verify that the structure function fit does indeed model the autocovariance at short lag times, which has not been previously shown. Furthermore, by proving and clearly stating that the sonic observations contain little noise, we can make the case that sonic observations are indeed a good baseline for comparison instead of simply assuming that they are. The text of the manuscript has been modified accordingly to make this clearer.

Within the updated manuscript, we have tested using a fixed lag time between 1 s and 10 s, as the reviewer suggested, and have updated the Fig. 6 and its discussion accordingly.

Specific comments:

- p. 4, l. 1: "the first in-depth analysis": it is true that you perform a very detailed analysis, but I would suggest to formulate this more cautiously because you investigate a rather short period (2 days) with mainly neutral to stable conditions (no characteristic convective conditions for sure)

The wording has been changed to 'we analyze the applicability' of the autocovariance method, which is better describes what is done in the paper.

- p. 4, l. 28: "two-day period between 26 March and 28 March" à can you give the exact dates? also in local time so that the reader knows how many nighttime/daytime periods are used?

Done, this information has been added.

- p. 6, l. 25: 10-m wind speed instead of surface wind speed?

Here, surface wind speed is meant, not the 10-m wind. It is clarified in the text that a no-slip condition at the surface is assumed.

- p. 8, eq. 2 / l. 18: it is a bit confusing that you use t as the time lag in eq. 2 and then say that $w(t)$ is a correlated variable, wherein t is again time

We agree, and we have changed the variable to simply w, indicating that it is an autocorrelated variable (implicitly implying it is a function of time).

- p. 11, eq.: the argument in the integral should be $M^*(t)$ and not $M(t)$

Here, M(t) itself is used directly to compute the integral time scale. By using M*(t), the integral time scale (as calculated using M(t) from the sonic observations) is significantly underestimated. However, for M(0) where uncorrelated noise is apparent, M*(0) is used. This is clarified in the updated manuscript. Fig. 1 below is provided to show the underestimate if M*(t) is used, due to misrepresentation of the autocovariance at longer lag times. The underestimate in comparison to the sonic anemometer for using M*(t) in the calculation of the integral time scale is shown in Fig. 2.

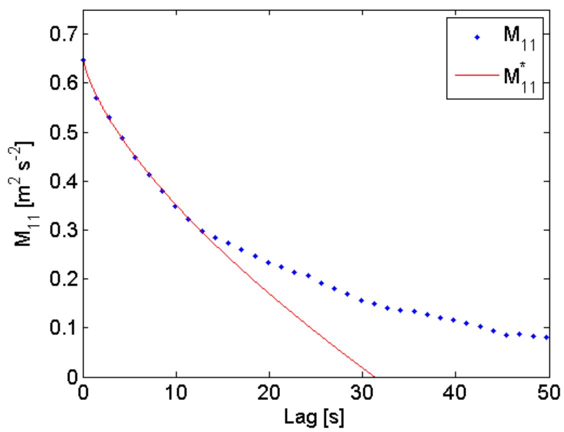

Fig. 1: M_{11} and M^*_{11} for a typical time period, showing underrepresentation of M_{11} by M^*_{11} at long lag times.

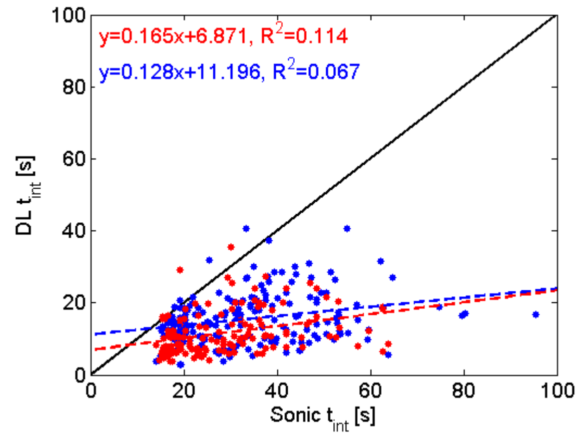

Fig. 2: Sonic and Doppler lidar calculated values of t_{int} using M^*_{11} instead of M_{11} in Eq. 9 in manuscript.

- p. 11, l. 11-14: leave out the two sentences ("Using this method …. compared to the small values of sigma_w"); I would not refer to the results section at this point and leave out these sentences

Done, these sentences have been removed from this section.

- p. 11, l. 22-23: slow advection also means that your time series sample is shorter in spatial dimension and less turbulent structures are contained within which makes statistics less reliable – this is of course related to larger integral time scale, as you say. Nevertheless, could the increased sampling error also be of importance? See also main comment (1).

See response to main comment (1). We have also added a sentence to the manuscript discussing that sampling errors are not a likely cause of the 'negative error'. Since the lidar beam is within a few metres of the sonic anemometer, sampling errors between the two instruments are small (see response to main comment 2, reviewer 1), and cannot explain the difference in the spectra shown in Fig. 4, d.

- p. 13, section 4.1: as said above (main comment 2), this subsection needs to be refined; first, describe only the averaging procedure for the sonic data. You can even consider to move this to a different section then, because it is more a method than a result. If you want to make a thorough analysis of the effect of averaging the sonic data, I would expect that you (1) show the averaging times which are used later in the study and (2) show the spectra. Did you also test a lowpass filter instead of averaging? I am also somewhat surprised that the averaging from 60 Hz to 1 Hz has more or less no effect on the variance while averaging to 0.1 Hz clearly eliminates relevant scales.

This can be much better understood when spectra are compared / shown.
This subsection has been moved to its own section before the results, as we agree with the reviewer that the paper will flow better if we keep its analysis separate from the rest of the results. We also have modified figure 5 to show one less example, and instead show the spectra for the raw and averaged sonic data.
The reason that we have chosen 1-s and 10-s averaging time is those are approximately typical for Doppler lidars (accumulation time typically varies from 0.5-2 s) and Raman/DIALs (~10 s accumulation), for which the autocovariance method is mostly used. The 1-sec averaging time is very similar to that of the OU DL, which takes data at ≈0.7 Hz. The LLNL WC measurement is an 0.67 s average (it simply only takes a vertical measurement every 4 s, the rest of the time it is performing a DBS scan), thus 1 s used in this analysis is between the averaging times for the two DLs used herein.

- in Fig. 10, it is obvious that the bias between sonic-derived and lidar-derived variances is smaller for WC than for OU DL, but that the scatter is much larger; this raises the question how you selected the examples in Figs. 8 + 9? are the examples for WC the best matches? Actually, I would have expected a better match of zero-lag autocovariance ("raw" variance) for OU DL examples shown in Fig. 8 than for the WC examples in Fig. 9, because of the higher sampling frequency of OU DL and due to the measurement height: you argue that deviations of OU DL from sonic-derived variance "are likely due to differing measurement volumes". If the measurement volume is the main reason, the difference should actually be smaller at higher altitudes because you can expect the dominant scales to increase with altitude. Thus, deviations should be smaller for the example in Fig. 8 (300 m) than in Fig. 9 (100 m). Are differences between "fitted" and "raw" variances significant compared to the difference to sonic values?

We clarify in the paper that the cases chosen were to represent different atmospheric stabilities and noise levels, where the effects of noise or volume averaging are clearly present. For the LLNL WC, this implicitly implies that some of the better matches were shown, otherwise these effects would not be as apparent. This is part of the reason why the zero-lag autocovariance for the WC look better than those for the OU DL.

In the updated manuscript, we discuss differences in accumulation times for the two lidar systems within the instrument description. Measurements from the WC are essentially averages over 0.67 s, while the OU DL averaged over 1.4 s. The longer accumulation time by the OU DL would limit its ability to resolve smaller-scale features. We also point out that the center of the OU DL range gate was a few metres offset from the sonic observation, whereas the WC range gates were centered exactly at the altitudes where the sonic observations are. Both of these factors may explain why the OU DL typically reports lower variance compared to the sonic anemometers. Other than these possible reasons, there is no clear explanation for these differences and we would not like to speculate on these differences.

McCaffrey et al. (2016) recently compared sonic anemometer observations on the opposing booms at the BAO tower, finding that the tower itself significantly impacts TKE measured on the booms around it. While this is especially true for waked observations, quantifiable systematic differences were found for other wind directions as well. Thus, it is difficult to compare sonic anemometer measurements on opposite sides of the tower with each other, as the turbulent structure is expected to be different.

- Fig. 9: can you mention values of Ri as in Fig. 8?
Done, this has been added to the caption.

- p. 17, l. 20. / Fig. 8: "flattening of the spectra": it is very hard to see this; the spectra are very noisy. Did you consider windowing for calculation of the spectra?
Great idea, we did not consider this previously. We have updated the figures and used a 5 point window on the spectra, which makes each of the spectra and the flattening much more visible.
- p. 22, l. 7-13: does this paragraph not belong to the previous subsection?
Correct, the paragraph (and the following one) belong in the previous section. They have been moved accordingly in the updated manuscript.
- p. 22. l. 14-15: it would be helpful to give the figure reference here (where can the slope of the best fit lines be seen?)
Done, a reference to the figure has been added.
- Fig. 11: you could draw a vertical line at 0.25 Ri to indicate the critical Richardson Number
Done, this has been added.

- p. 22, l. 26 / l. 30: "When conditions are stable […], fitting is less clear" – "applying the extrapolation technique during stable conditions generally improves the estimates" à isn't that a contradiction?
We have modified the discussion here, to ensure there is no contradiction. Instead of making a blanket statement about using the extrapolation technique to improve estimates of variance, we suggest that it only be used if the SNR is below a threshold.

- p. 24, l.9: You could add that especially unstable / convective conditions were not Present
Done, added phrase to emphasize this.

References:

McCaffrey, K., et al., Identification of tower wake distortions using sonic anemometer and lidar measurements. Atmos. Meas. Tech. Discuss., in review, doi:10.5194/amt-2016-179, 2016.