Reply to referee 1 (RC1)

We thank the reviewer for their comments and suggestions that have helped improve the manuscript, please find our replies below. Please note that the line numbers refer to the revised manuscript (with track changes).

- The parts of the abstract highlighted by the referee have been rewritten (lines 10 -15).
- Two references have now been included (line 29-30).
- Brown et al. in fact speculate that inaccuracies in aerodynamic modelling could account for a certain portion of the errors in the one-to-one comparisons. Please refer to the following line in section 6 pg. 17 of the article: “However, the simulated flapwise blade-root DELs and tower-base fore-aft DELs showed significant median biases of up to ∼40% overprediction, which the authors speculate could be a result of inaccurate aerodynamic modelling in sheared conditions (note this shortcoming is being addressed currently by NREL in the ongoing development of OpenFAST) and omission of certain control features of the proprietary field controller, combined with possible errors in the simulated inflow wind fluctuations, and other unidentified errors.” As per the referee’s suggestion, we have added a note about the work of Doubrawa et al (line 58-60).
- Corrected (Table 1).
- The font size in the figures has been made more consistent.
- The sentence has been made clearer. By “three decades” we meant 10^3 or 3 decades on a logarithmic scale (line 129).
- Corrected (line 166).
- The legend has been relocated.
- Corrected (line 223).
- The sentence has been reworded (line 269).
- We used the word “however” because our methodology differs from that used by Sathe et al. (2013) who analyzed loads from aero-elastic simulations only. The results are indeed similar, more precisely please compare Figures 9 and 10 of Sathe et al. to Figures 6 (a) and 7 (a) from our work. Since the sentence is after the discussion of the results related to the effects of atmospheric stability, we believe that the placement is appropriate.

Reply to referee 2 (RC2)

We greatly appreciate the referee's in-depth review. We believe the points raised by them have helped improve the manuscript. Please find our replies below:

Specific comments:

1. A more precise reference has been added (line 27).
2. A correction has been made (line 44).
3. Sentence revised for clarity (line 63).
4. The reference has been corrected (line 61).
5. A sentence in the second last paragraph of the introduction has been added (line 71 – 73).
6. No, we mean a turbine operating in the real world, i.e., we show the relationship between fatigue loads and atmospheric stability using load and wind measurements. Holstlag et al. (2016) provides an analysis of the NREL 5 MW via simulations.
7. Changed (line 87).
8. The paragraph has been made more concise (line 92 – 96).
9. Yes, we believe 5 Hz is more than sufficient as this frequency lies well within the initial sub-range of the turbulence spectra. Indeed, as seen in Figure 3, not much energy is left in the spectra even at 2.5 Hz. Thus, we can be confident that measurements capture enough of the spectra to obtain a good quality fit with the turbulence model. From the perspective of loads, turbulent fluctuations higher than even 1 Hz are not deemed to be important as the turbine behaves as a low-pass filter. Moreover, the structural eigen frequencies are much greater than 5 Hz.
10. A table on data availability has now been added (Table 2).
11. Averaging period has been added (line 98).
12. The sentence has been modified.
13. We acknowledge that the distribution of stability at 155 m and 241 m is generally not the same. But for a given location we would expect it to have the same trends, i.e., if unstable conditions are common at 241 m, then they should also be relatively frequent at 155 m. So, our reasoning that the two datasets have a different seasonality, which causes the disagreement in the relative distribution of stability, still holds true.
14. We agree with the referee that using the power law profile to fit measurements from only three heights is a simplification. However, this method is often used in industry (and many older academic studies). Thus, it is important to highlight the drawbacks of this method, most notably by the fatigue load comparisons presented for the turbine blades in section 3.2. With regards to the quality of the vertical profile fits, the lowest

value of R^2 from the 36 bins was 0.92. Thus, generally the wind speed profile up to hub-height fits the power law well.

15. Anonymization is just a rescaling operation which is the same for all the data in each plot. Hence, it preserves the qualitative difference between data in the figure. To comment on the quantitative aspects, we have attempted to do so via the text in section 3.1 and the z-scores in section 3.2. Anonymization is of course less than ideal as it does not show the absolute magnitude of the loads leading to some confusion as highlighted by the referee's comments no. 20 and 21.
16. A figure showing the measured turbulence intensity for each bin is now added to the manuscript.
17. The figure caption has been updated.
18. For reasons of confidentiality, we cannot mention the precise rated wind speed of the turbine. We have changed the sentence to give an estimation of it.
19. The sentence has been modified.
20. The edgewise fatigue damage does appear to be lower in unstable conditions however the difference is very small. Although this is difficult to show in absolute terms due to the anonymization, please consider that the difference between the maximum and minimum edgewise loads at 21 m/s and 5 m/s respectively, is four times lower than the minimum edgewise loads (since the y-axis starts from 4). Edgewise loads are typically in order of 10 MNm. Thus, the difference between the max and min edgewise loads is in the order of 1MNm. Unstable conditions apparently show edgewise loads 0.3 units lower. Thus, this difference is in the magnitude of 0.1MNm which we believe is not very significant.
21. The referee is correct in pointing out that the mean measured loads appear to be higher than the simulated loads but when including the variance of the individual 10-minute measurements or seeds in case of the simulations, we see that the difference is less than 3 sigma. Thus, in our opinion, the difference is better explained by random errors rather than any systematic error in the modelling framework.

Technical corrections:

1. Equation (2) has been corrected
2. A unit has been added.
3. No, we do not, as the fit is derived using Eq (6). Equation (7) relates the 1D spectra to the spectral tensor. Indeed, the 1D spectra are precomputed into a look-up table, so when fitting the model to the measured 1D spectra, only Eq (6) must be evaluated.
4. Fixed as also suggested by referee 1.
5. Sentence has been modified.