Question about assumed homogeneity; stable almost neutral
Anonymous Referee #2

The paper presents a validation of WRF LES model on three canonical flow cases: neutral, stable, and unstable. The motivation, structure, content, graphical presentation, and the references, are all excellent.

The high-quality data for the validation is collected at Østerild test centre, from 5 heights between 7 and 240 metres.

The three flow cases were defined using the Monin-Obukhov stability at 37 meters above the ground. The surfaces heat fluxes and/or surface temperature tendency, and the roughness length, resulting in the observed variables, were used to set-up the WRF model. This approach is adequate, but the assumption of the homogeneous surface should be validated to ensure that the flow characteristics are not too much affected by the fetch distance for different heights and wind speeds. It is possible that the apparent roughness length would be different if other heights than 37 metres would be used to derive the necessary parameters.

Given that a lot of effort has been invested into estimations of the surface roughness using satellite and lidar data, especially around Østerild, one could ask why you haven’t used the available roughness data for these WRF simulations. Are you suggesting that the roughness length is a function of the flow, and not the other way around?

The model's capability to accurately simulate the three flow cases is nicely analyzed in terms of the mean properties, and all stress components. The agreement between the LES model and real data is quite good. In particular the velocity spectra are fantastic, in the resolved part of the spectrum of course.

All three cases are shear-driven which, especially for the stable case, somewhat limits the applicability of the results. You have proposed a few directions for further work and we can look forward to it being performed and presented.

Specific questions:

1. What was the period of the data collection, i.e. how much data is actually used in the
aggregation for U, N, and S? From Figure 1 it seems that about 6 months of data is used, however from L274 it follows that there is 3000 seconds of data. Please clarify.

2. Figures 3-5: would it be possible to estimate the Ri number for these cases? It could be useful for comparisons when more of the flow cases will be constructed in the time to come.

3. Figure 6: it would be easier to compare the instantaneous flow representations if the color scales were the same.

4. P12L203: Here the RMSE is introduced, but it is not immediately clear if this is the statistics derived from 5 values (5 heights, and using the whole-period-average values), or? Please clarify.

5. P12L214: Please elaborate/comment on the possible reasons why the neutral simulation differs most from the observations, i.e. the stable and unstable simulations match the observations better.

6. Figures 9-11 (also elsewhere but perhaps easiest to discuss here): the differences between N and S cases are very small, it requires quite an effort to find any significant differences. Have you considered finding "more stable" flow cases, or would that be difficult at this location?

7. P16L262-267, Figure 12: Please expand the discussion of the TKE profiles. For example, the neutral simulation seems to "compress" the TKE profile towards the surface, evident by the "nose" of the profile being quite lower than 37 metres, which is the height of the maximum observed value.

8. P16L272: diving --> dividing

9. P21L339-340: there is some confusion regarding the inversion strength, and the adiabatic lapse rate for dry atmosphere. The capping inversion and the adiabatic lapse rate should have the opposite signs, and are unrelated in this context. What was the purpose of the statement in L339-340?