Review of *"Model sensitivity across scales: a case study of simulating an offshore low-level jet"* by Patrick Hawbecker, William Lassman, Timothy W. Juliano, Branko Kosović, and Sue Ellen Haupt.

The manuscript provided by the authors addresses the application of different sea surface temperature (SST) data sources to model specific events across different modelling scales and validate them against floating lidar measurements. The objective of the presented study is to improve the understanding of the sensitivity of simulation frameworks across different scales on SST datasets and find the best performing dataset. Simulations are carried out in the Weather Research and Forecasting model (WRF) using five nested domains with increasing horizontal resolution, ranging from 6250 m to 10 m, thus covering the mesoscale, grey zone of turbulence and microscale. Specifically, the study deals with one offshore Low-Level Jet event observed by a floating lidar device off the coast of New Jersey in the United States comparing the measured physical features, such as vertical wind shear, jet core height and speed to the characteristics obtained by the seven simulation members driven by different SST datasets.

The authors found that the best-performing SST dataset throughout all domains is the CMC SST dataset. For some other datasets large discrepancies between performance in microscale and mesoscale domains is observed. The OSPO dataset for example performs second best in the mesoscale while showing the worst performance in microscale. The opposite is true for the OSTIA dataset, as it secures a third place in the microscale, while performing worst in the mesoscale. Further, the authors acknowledge, that these results are obtained from observations of one specific event at one location and also underlie the subjective weighting of different metrics to generate a ranking. Thus, the authors claim that while the process of determining good performing setups is well-suited, the overall ranking is not generalizable for other situations.

The article provides valuable inputs for the modelling community, but could benefit a lot from a broader discussion about the generalizability of it's findings, i.e. how the findings could be transferred for a broader audience facing similar challenges with different meteorological situations. Also, the mainly time-series based visualization of results sometimes is hard to follow and to interpret. Here, a more a more concise and systematically organized presentation would enhance their clarity and readability.

General comments

1. The introduction could benefit from short description about the formation of LLJs, i.e. why the temperature difference between sea surface and air plays a major role in their development.

2. When describing the considered LLJ event in Section 2.3, some meteorological information about the presented case, such as e.g. the wind direction, is missing. A more elaborate information about the event and

why this specific case is of interest would be helpful in following the story of the manuscript.

3. Instead of relying on time series representations that much in the results, maybe a depiction of e.g. the correlation between the meteorological features would make it more easy to spot good or bad performing setups quickly. The way the results are presented right now is quite repetitive and differences between the different setups are hard to interpret.

4. In Section 4.4 you briefly touch the process of selecting the "best" performing setup to run the LES domains. While you recognize, that this process is subjective and depends on the region and event of interest, maybe you could elaborate a bit more on the process of how you chose which metrics are of importance and how you weigh them when reaching the conclusions. This more transparent approach would also be of great help in the Methods section to understand the reasoning behind the entire process.

5. In my opinion the manuscript would benefit of a separation of Results and Discussion. Right now, some discussion of the results is already performed in the Section 3, while some of it is also present in Section 5, making it sometimes hard to follow the storyline of your paper.

Minor comments

1. L. 64: Here an "area" is missing after "rotor swept"

2. L. 103: I do not fully understand why the highest resolved dataset (MUR) should not be smoother than lower granularity datasets such as OSTIA or GOES-16. In my mind the highest granularity dataset should provide the smoothest gradient. Could you elaborate why that is a contradiction?

3. L. 111 Here, it would be helpful to include the information, that the buoys are also used to gather the Temperature profiles and SST.

4. L. 163: If I am not mistaken it should read "LES turbulence *closure* techniques" not "*closer*"

5. L. 164: You mention the use of LES turbulence closure techniques is questionable for these resolutions. Could you maybe elaborate on why that is and why you still chose to use this parametrization over mesoscale PBL schemes which are "even more questionable"?

6. Figure 4: Consider changing the numbers to markers in the Figure, as this depiction looks rather confusing than helpful for me. Also, I'm not quite sure, what the spread here depicts. Is it just the difference between maximum and minimum value for the different SST datasets?

7. L. 193: "Wind speed *maximum*" instead of "*maxima*"

8. L. 200: What is here meant by "shifting" the data? Is this just a temporal shift? Regarding this point it would also be interesting whether grid or spectral nudging is used when driving the simulations. Could you please elaborate here?

9. L. 214: When referring to shear, are you talking about the bulk shear ($\frac{z_{\mathrm{up}}-z_{\mathrm{low}}}{\Delta z}$) between lower tip and hub height, or the average shear across that region ($\sum_i \frac{u(z_{i+1})-u(z_i)}{z_{i+1}-z_i}$)?

10. L. 226: For consistency, please align the depiction of means. In Figure 8, they are shown as overlines in the Figure titles.

11. Figure 9: In all other figures, domains in the legend are not abbreviated. Consider aligning for consistency. The same is true for Figure A2.

12. Line 291: Are bias and RMSE calculated in reference to the Observations here?

13. Table 2: For brevity you could consider summarizing D01 and D02 as mesoscale, D03 as grey zone and D04 and D05 as LES domains. This would make the table more accessible as the results for both mesoscale and microscale domains, respectively, are the same anyhow.

14. L. 368: Is this a new research question? Consider already adding this part to your objective statement in the Introduction.