Original comments; Author responses

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.

We would like to thank the reviewer for their summary and suggestions for improvements to the manuscript. The suggestions for improvements above are itemized in the general comments and are addressed below with the exception of the generalizability of the study. We appreciate the request to address the generalizability of these findings and have added the following text to the manuscript:
Line 69 of original manuscript:
We note that the study considers a single case study for a specific topic; thus, it is unclear whether the findings generalize to other cases and atmospheric phenomena. However, we explore fundamental differences in mesoscale and microscale simulation techniques that are generally applicable for other atmospheric studies.
Line 375 of original manuscript:
These fundamental differences emphasize that studies should generally use caution when assuming that mesoscale sensitivities will directly translate to the microscale when simulating atmospheric phenomena (such as low-level jets) that have known dependencies on model grid spacing and turbulence closure techniques.

Original comments; Author responses

Line 383 of the original manuscript:
Additionally, similar studies for additional cases, different atmospheric phenomena, and different parametric sensitivities will be required in future studies to determine when and where mesoscale sensitivity can directly translate to microscale sensitivity.

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.
   We appreciate the reviewer's suggestion and have added the following text to the manuscript:
   Line 47 of the original manuscript:
   Offshore low-level jets can form when relatively warmer air advects over colder water temperatures generating stable conditions and frictional decoupling of the winds aloft (see De Jong et al., (2024) for more information on offshore low-level jet formation mechanisms in this region).
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.
   This is a good suggestion and the following text has been added to the manuscript:
   Line 114 of the original manuscript:
   The dominant wind direction in this case was from the South indicating the possibility of warm air advection as is commonly seen in the region (De Jong et al., 2024).
   Line 120:
   This case was selected due to the clear and consistent LLJ signal and the apparent dependence on the air-sea temperature gradient.
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.
   We appreciate the suggestion of the reviewer but do not agree that the analysis is redundant, and in fact, by comparing many variables over time, one can compare against variables/Figures as is done in the manuscript. It is discouraging to hear that the reviewer finds the analysis repetitive but we do not feel it is helpful to change the presentation of the results for the sake of variety.
   Because this is a case study and we have insufficient data points to draw meaningful correlations, we find that the additional analysis of correlation does not reveal better performing models; and thus, do not include analysis that does not show significance. In this study, however, we considered the quantitative metrics of bias and error..
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.
We thank the reviewer for recognizing how this section relies on subjective interpretation and have added the following text to clarify our ranking:
Line 307 of the original manuscript:
Ranking the performance is highly dependent on the specific feature being studied. For this study, we consider performance of the dynamic variables (low-level shear, hub-height wind speed, and REWS) and forcing conditions (2-m Temperature, SST, and ΔT) to rank the mesoscale setup performance. Considering these variables and weighing them equally, we rank the mesoscale SST dataset performance from best to worst as follows: CMC, OSPO, GOES-16, Default SST, NAVO, MUR, OSTIA (Table 2).

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
   We appreciate the comment from the reviewer on how to make the storyline more coherent. We are not sure where in Section 3 (Model Setup) the reviewer refers to results being discussed. If you refer to initial testing of different reanalysis dataset (ERA5), that mention was merely an explanation of why we chose MERRA-2 as the reanalysis rather than a “result”. We also discuss in Section 3 how a WRF pre-processing bug impacted our study and refer the reader to an appendix on the issue and its impact. We have split the summary and discussion sections into separate sections (Sections 5 and 6, respectively).

Minor Comments:

1. L. 64: Here an ”area” is missing after ”rotor swept”
   Thank you for finding this error; it has been corrected in the revised manuscript.
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?
   We thank the reviewer for the question but would like to suggest that the highest granularity dataset does not necessarily imply the sharpest gradients. We agree that the MUR dataset does appear to resolve less energy at small scales than other datasets, but describing how each of these products is generated is beyond the scope of this paper.
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.
   We have adjusted the sentence on line 107 by adding the following text to the manuscript to clarify this:
   Line 107 of the original manuscript:
   These buoys, named E05 and E06 (open and filled X in Figure1, respectively), contain vertically scanning lidars, ocean and wave sensors, and a small meteorological mast recording atmospheric variables such as temperature and pressure.

Original comments; Author responses

4. L. 163: If I am not mistaken it should read ”LES turbulence closure techniques” not ”closer”
   Thank you for finding this error; it has been corrected in the revised manuscript.
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”?
   We thank the reviewer for their question and have added the following text to the manuscript:
   Line 157 of the original manuscript:
   LES turbulence closure techniques are used in this study within this domain although its appropriateness is questionable due to the fact that the largest energy-containing eddies are not fully resolved with this grid spacing. The other option is to use a planetary boundary layer scheme at this resolution, but the assumptions of horizontal homogeneity and that all energy-containing eddies are unresolved renders the applicability of such parameterizations at 250 m grid spacing are yet 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?
   We thank the reviewer for the suggestion. We have tried using symbols for this figure but the same issue crops up where all or most of the symbols are on top of one another and in the end, the figure is not more clear. Considering that, we found the numbers more useful as there is less question about what “1” represents (domain-1); whereas if a circle were used, it is not as obvious and readers are left mapping the caption to the figure. Instead, the figure highlights that other than domain 1, the majority of the domains have similar representations of sea surface temperature in the domains. The spread quantifies exactly how different the average SST is between domain 1 and 5 as noted in the caption of Figure 4.
7. L. 193: ”Wind speed maximum” instead of ”maxima”
   Thank you for this suggestion. The sentence has been reworded to be, “During the period of interest, maximum wind speeds are observed between 80 and 150 m (Figure 5a)
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?
   We see how this is unclear as originally stated. The sentence has been adjusted to read, “When shifting the observations in time to better match with simulated results…”
9. L. 214: When referring to shear, are you talking about the bulk shear between lower tip and hub height, or the average shear across that region?
   On line 185-186 of the original manuscript we define what we mean by low-level shear but have added the term “bulk shear” to indicate that it is indeed the former. The sentence now reads, “We define low-level shear as the bulk shear between the bottom of the rotor swept area (38~m) and hub-height”

Original comments; Author responses

10. L. 226: For consistency, please align the depiction of means. In Figure 8, they are shown as overlines in the Figure titles.
    We thank the reviewer for the suggestion but are unsure as to what needs to be adjusted. Line 220 notes that angle brackets represent ensemble average, while Line 229 explains that the overbar represents time average. We do not see an instance in which this is misused and would appreciate it if the reviewer clarified where this is done so we can change it.
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.
    We thank the reviewer for catching this inconsistency. The figures have been updated in the revised manuscript.
12. Line 291: Are bias and RMSE calculated in reference to the Observations here?
    Yes. To clarify this, the sentence now reads, “Considering root mean square error (RMSE) and bias with respect to observations for each setup on the mesoscale and microscale domains…”
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.
    We thank the reviewer for the suggestion and have condensed the table.
14. L. 368: Is this a new research question? Consider already adding this part to your objective statement in the Introduction.
    It is unclear what statement the reviewer is referring to with this comment. We are assuming it is about the sentence reading, “This finding suggests that although we can try to set up our LES simulations to have the best chance of success, the differences between the mesoscale and microscale numerical methods and model setup are large enough that one of the best performers on the mesoscale may end up being the worst performer on the microscale.”
    We believe that this is addressing one of the central research questions of the study of, “can we expect the microscale simulation driven by the best performing mesoscale setup to produce the best microscale result?”