Original comments; Author responses

Below are point-by-point responses to the editor and both reviewers.

Editor Comments

The two reviewers noted substantial improvement of the manuscript and recommend publication after minor revision. Please read their comments carefully and address them adequately. I will perform the final check after your revision.

In addition, based on my own reading, I have the following observations:

1. Methods content should be moved out of the Results section
   Some material currently in Section 4.3 (in particular, Equations 1–3 and the description of ensemble metrics) belongs in Section 2 (“Methods”). The Results section should present only the outcomes and their interpretation, not methodological explanations. Moving these elements to Section 2 will significantly improve clarity. (By analogy: the recipe is given before the cake is baked, not after.)
   We understand this correction and the equations and descriptions have been moved to the methods section.

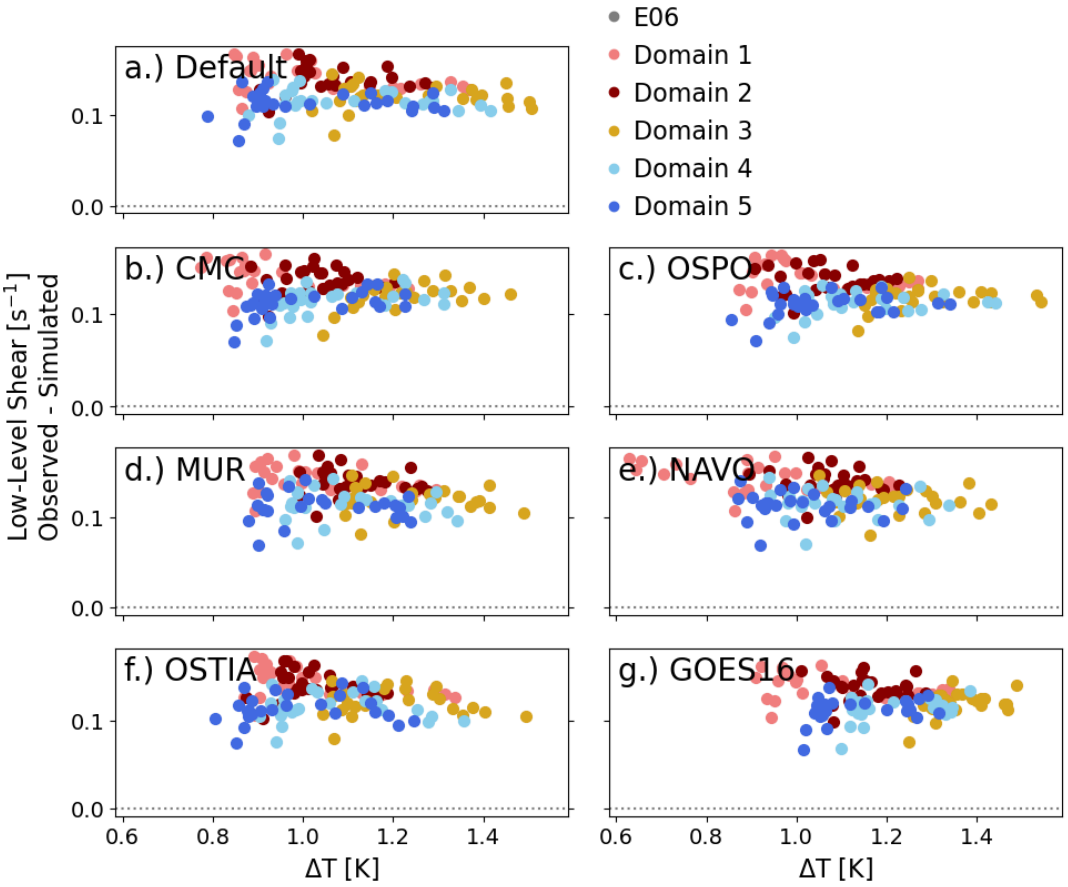
2. Missing units in Figures 9, 10, and 11
   Figures 9, 10, and 11 appear to be missing units on the y-axes (temperature, wind speed, shear, etc.). Please revise the axis labels accordingly. The font size of the figures can be slightly reduced if needed to accommodate units.
   We thank the editor for catching this oversight. Units have been added to all figures with missing units.

3. Clarification needed for Figure A1
   In Figure A1, the right-hand panels (domain 2) show aliasing artefacts that could mislead readers. In addition, the y-axes of these right-hand panels are not labelled, implying they are identical to the left-hand panels (domain 1). Please confirm whether that is correct and ensure the figure is unambiguous. I also note that “domain 1” and “domain 2” terminology has already been used in Figure 3. Are these the same domain? Consistency and clarity here are important.
   We recognize now that the Appendix was poorly describing the problem and have modified the text appropriately. The Appendix addresses the artifacts within the figures and the workarounds we made to approach it correctly. They emphasize that if using WRF at high resolution and WPS is used in single precision, you will have errors within your simulations. This issue was reported to the WRF/WPS developers and corrections have been added to the official release of WPS version 4.5.
   The labels in the y-axis have been added and the names have been changed to Domains 1A and 2A to further emphasize that the domains are different from the main sections of the paper.

Original comments; Author responses

4. Possible relocation of Figure 12
   Figure 12 quantifies air–sea temperature differences across SST datasets and domains. Because this information characterizes the inputs before the simulations are run, it may be better placed in Section 2 or Section 3 rather than in the Results section.
   This figure includes both inputs (SST) and outputs from the model (2-m T and delta T), thus it seems best to leave it after the model calculations are explained. We have modified the figure caption to clarify this point.
   "Same as Figure 7 but for 2-m temperature, SST, and ΔT. While SST is defined from the input SST datasets, both 2~m temperature and the resulting ΔT are predicted within the model."
5. Clarification and strengthening of the quantitative comparison (reviewers' main remaining concern)
   Reviewer 1 notes that the manuscript would benefit from a clearer and more complete presentation of the quantitative results, particularly those related to the correlation analysis. Your response mentions high p-values, but does not specify which variables were correlated with which others. As a result, it is difficult to interpret Table 2 or understand the attempted correlation analysis.
   We see that we had neglected to clarify what was being compared in the correlation analysis. As with the other metrics in the study, the p-values from the correlations shown in Table 2 compares model vs. observations. Variables considered are noted in the top of Table 2: low-level shear, hub-height wind speed, and REWS.

   There may have been a misunderstanding around the word "correlation." Reviewer 2 was referring to the correlation between meteorological features. Specifically, they suggested a concise quantitative way to see how environmental inputs relate to model performance. Examples would include:
   Correlation (or simply quantitative metric) that shows the link between ΔT (air–sea temperature difference) and low-level shear error, between SST bias and hub-height wind speed bias, etc... Such relationships can be presented descriptively (e.g., correlation coefficients, scatter plots, etc...), without relying on statistical significance tests. The aim is interpretability: helping the reader see which setups tend to perform better and why.
   Unfortunately, analyzing the relationships of the suggested variable combinations (ΔT vs low-level shear error; SST-bias vs. hub-height wind speed, etc.) do not make things more interpretable for the limited number of hours modeled. Considering ΔT vs low-level shear difference (shown below), we see no notable relationship and the significance of the resulting correlation coefficients is above 0.1 for 75% of the datasets.



For the correlations that are of significance (p-value < 0.1) between ΔT vs low-level shear difference (table below), we see that they are either roughly -0.4 or +0.4 (and average to nearly zero). Unfortunately, this does not add additional insight to the study.

Setup-domain	Corr.	Sig.
Default-d01	-0.402	0.046
CMC-d01	-0.423	0.035
CMC-d03	0.446	0.026
CMC-d04	0.438	0.029
CMC-d05	0.446	0.026
NAVO-d01	-0.456	0.022
OSTIA-d01	-0.425	0.034
GOES16-d01	-0.340	0.096
GOES16-d03	0.406	0.044

SST-bias is effectively a constant value, so correlation between that and any other variable is undefined.

We have analyzed the data in a quantitative manner within this study and have not

included correlations because there simply is not enough data to draw conclusions on these relationships. We strongly believe that correlation between any variables without analysis of the statistical significance does not belong in a scientific journal. The statistical significance determines whether the results are meaningful.

More generally, the paper would be clearer if its structure more explicitly reflected the two-step logic of the study: (1) How the choice of SST dataset and domain affects the modelled meteorological state (this is well presented), and (2) How these differences translate into agreement or disagreement with the lidar observations (this is the part that may have been insufficiently quantified). I think that addressing even better the quantitative link between (1) and (2) would fully address the remaining reviewer concerns.

We agree with the reviewer that step (1) is well presented in the paper, but believe that step (2) is also well presented. The lidar observations are included in every metric shown with the exception of surface sensible heat flux in Figure 12c. The domain effects are highlighted in several areas in which the difference between the boundary layer parameterization and subgrid turbulence model drastically change the simulated LLJ characteristics. We do not heavily focus on the impacts of SST directly as it is not within the scope of this paper. To clarify the paper's intentions, we have adjusted the introduction to include the following paragraph:

"*In this study, model sensitivity of an offshore low-level jet (LLJ) to sea-surface temperature (SST) is analyzed across both the mesoscale and microscale.*
The goal is to assess, on both the mesoscale and microscale, the sensitivity of LLJ characteristics to SST and model performance when compared to observations in order to determine whether the assumptions above are indeed valid."

Reviewer 1

I would like to thank the authors for carefully revising their manuscript. Once the point raised by the second reviewer and the editor on the extension of the presentation of quantitative results has been properly addressed by the authors it can be accepted for publication. I detected only one typo in the revised manuscript. In line 351 "weighing" should be replaced by "weighting".

We thank the reviewer for their comment and in catching this spelling error. It has been replaced in the manuscript.

Reviewer 2

The authors revised their manuscript and implemented helpful changes to improve the manuscripts quality. Below some further comments on the revised version of the manuscript are collected. To proceed I would recommend the paper to be accepted, subject to minor

revisions.

We thank the reviewer for their comments and suggestions.

1. L.9: There seems to be a typo: ”scehme” instead of ”scheme”
   We thank the reviewer for catching this error and have updated it in the manuscript.
2. L.131: Thank you for clarifying the used LLJ detection algorithm. In my opinion, it would be helpful to the reader to also directly describe the used definition, since it — as described in Debnath et al. (2021) — uses a combination of shear and fall-off criteria. This might be useful information for the reader down the line, when you analyse the low-level shear between different simulation set-ups.
   We have added a short description of the method to the manuscript but want to note that while the observed LLJ was detected using the Debnath et al. (2021) definition, the simulated LLJs are defined strictly by the height of maximum wind speed. Once detected in observations, we no longer use that algorithm.
   Added text: *“This algorithm detects a low-level jet based on meeting three criteria: (1) the maximum wind speed is not at the first or last lidar level, (2) the level of shear between the lowest lidar level and maximum wind speed is above 0.035~s^{-1}, and (3) the wind speed drop off between the maximum wind speed and top lidar measurement is greater than 1.5~m~s^{-1} and the drop-off is more than 10\% of the maximum wind speed.”*
   Similarly, it would be helpful, to see what values for shear and fall-off were detected in this specific event (e.g. in L. 216ff).
   The values of low-level shear for the observed LLJ can be found in all panels of Figure 7. We have not considered drop-off as a metric to compare within this study because it is limited to 200 m (the maximum height of the observations). We do not consider the drop-off value to be within the scope of the paper.
3. Fig. 7/ Fig.8: From the time series data in Fig. 7 and the profiles in Fig. 8, it is seen that the low-level shear, as well as the fall-off is considerably smaller for the mesoscale domains. Could you elaborate on whether the LLJ definition you applied, detects the LLJ throughout all domains and all different set-ups.
   This is true that the mesoscale domain has a higher jet nose height and so, when limiting the view to 200 m, the drop-off is much lower than the microscale domains and observations. For the simulations, the low-level jet is defined strictly by the height of maximum wind speed. An LLJ detection algorithm was used to find an event to simulate in this study, but the analysis of the detection algorithm performance is not in the scope of this paper.
4. L.285/Fig. 9: For both hub height wind speed and REWS, EMEs larger than 1 ms−1 occur at times. Given that you already calculated the REWS, would it be possible to elaborate on how these differences in wind speed translate to differences in possible power production, as power changes with the cube of the velocity. I see, that your specific case shows wind speeds that are probably above rated wind speed for the turbine sizes you assume. This actually makes it a two-part comment: a) How do the EMEs and Spreads translate to lower wind speeds and b) how large is their effect on

an exemplary turbine's power production?

We appreciate the suggestion to analyze potential power production impacts and have included the following text in the manuscript:

"Note that the wind speeds modeled are in the rated portion of most wind turbines. For reference, if we were in the cubic portion of the power curve, over-prediction of wind speeds by this amount would result in over-predictions of energy production during this period by between 3–16% for the mesoscale domains and between 15–27% on domain 3 and the LES domains (assuming wind speeds are below rated wind speed and above the cut-in speed, a performance coefficient of 0.4, and an average air density of 1.225 kg m^3)."