Referee 1

Referee comments appear in black and author responses appear in blue.

This paper is well-written, generally well designed and deserving of publication. I noticed that there have been previous review rounds, but I was not involved in those. However, I would like to raise some points that, in my opinion, should be addressed to enhance the paper's quality.

We thank the new reviewer for assisting with the review process and providing feedback to enhance this research.

1. Firstly, it would be beneficial to provide more information about the model's grid setup. The distance between the outer edges of the small domain and the outer edges of the large domain is rather small, and I wonder if the lateral nesting zone of the small domain is not in the spatial spin-up zone of the larger domain. Even though the model setup relies on a previous study, it would be good to clarify this.
While it appears that there is not much distance between the inner and outer domain boundaries, this is a very large domain. There are 20 cells between the outer and inner domain boundaries, which allow 4 waves to be resolved before entering the inner-most domain. We acknowledge the rule of thumb that the inner domain boundaries should be at a distance of roughly 1/3 of the outer domain's total extent, but our setup is 4 times the minimum requirement.
We have modified the text on lines 108-110 to: "Two nested domains comprise 6 km and 2 km horizontal resolutions (Pronk et al. 2022; Xia et al. 2022; Bodini et al. 2023; Redfern et al. 2023), respectively, *and the inner nest begins 20 grid cells into the parent domain* (Figure 1)."

2. I hold the opinion that using (bias-corrected) Root Mean Square (RMS) and correlation as metrics for evaluating model performance may not be the most suitable choice. Even though spectral nudging is applied in the outer domain, the flow in the inner domain can evolve freely. A small deviation in a weather system's position or timing could result in a double penalty effect. Therefore, for this type of studies, it is crucial to focus on getting the statistics right, rather than precisely timing weather systems. In my view, bias and quantification of distribution metrics are more relevant for such studies. I recommend a brief discussion of this issue when describing the evaluation metrics.

We believe that reporting more metrics than fewer metrics benefits a wider readership. We have added a discussion of how these validation metrics benefit model validation on new lines 237-240:

"Each of these metrics provides different insight into the performance of the model. For instance, the correlation coefficient illuminates how well the model captures the timing of weather systems and diurnal variability. EMD emphasizes the difference between distributions but not the timing. Bias captures the difference between measured and modeled values. Finally, cRMSE describes the random component of error."

3. I would argue that there should be only one definition for external loss. The total loss can be accurately represented by equation 11, as it is defined based on the reference no-wind farm simulation. Similarly, the internal loss for only the Vineyard Wind farm is correctly calculated, as it also involves dividing by the no-wind farm simulation.   However, I have reservations about the correctness of equation 9. Your reference here is the power from the Vineyard Wind, rather than the no-wind farm simulation. The definition of external losses should be the difference between the run that includes all wind farms (combines internal and external loss) and the run with only the Vineyard Wind Farm (internal loss), divided by the power from the no-wind farm simulation (which is the reference). This is equivalent with Equation 12.

While we appreciate the reviewer's request for a streamlined definition for external loss, our experience in presenting this work in several industry and academic venues suggests that different communities request different definitions of external loss. Each of these definitions can be justified, and so we feel that it is important to include multiple measures. We have received requests for the loss definition as defined in the former equation 9 (new equation 15). We thus choose to present multiple definitions, clearly defined, so that readers can use the definition most suited to their interests. We believe it is helpful to include as an alternative method for calculating the external power losses.

4. The averaging of percentual power loss might be incorrect: I would argue that the percentual power losses, is the total power loss over a period considered (could be total year, or the January 0-1UTC periods), divided by the power of the reference situation (no wind farm) over the same period. However you describe that first you calculate the percentage loss over an hourly window and then these percentages are averaged. This is not the same.

Yes, we agree that using the percentage losses over an hourly window would not be the same. We choose to average over an hourly window first before all subsequent calculations to reduce the effects of numerical noise.

5. In many climatological studies, the Perkin Skill Score is utilized to compare distributions from climate models for various variables. Both the Earth Mover's Distance (EMD) and the Perkin Skill Score are closely related metrics. It would be valuable to include a brief description of how those two compare. Personally, I find the Perkin Skill Score very intuitive because it is dimensionless and assesses the overlap between distributions.
We have added the following text to the methods section on lines 459-478:

We calculate profiles of the Perkins Skill score (PSS) (Perkins et al., 2007) between NWF and lidar wind speeds. Wind speeds are considered at 20 m height intervals from 20 m to 200 m. Each wind speed timeseries is subset by all timestamps with unstable, stable, and neutral stratification, respectively. After subsetting, timestamps where lidar observations return NaN are removed from both lidar and NWF timeseries. At each height, the probability distribution functions of wind speeds are binned at 0.2 m s–1 intervals and normalized such that the frequencies add to unity. The minimum frequency between modeled and observed values for each bin is stored, and the resulting stored values are summed to calculate the score (Eq. 12):

$$PSS = \sum_{i=1}^{n} \min\left(C_{NWF}(z), C_{lidar}(z)\right) \tag{12}$$

where n is the number of bins, C is the count of normalized values in a bin, and z is the height. A PSS of 1.0 suggests perfect overlap of the two distributions.

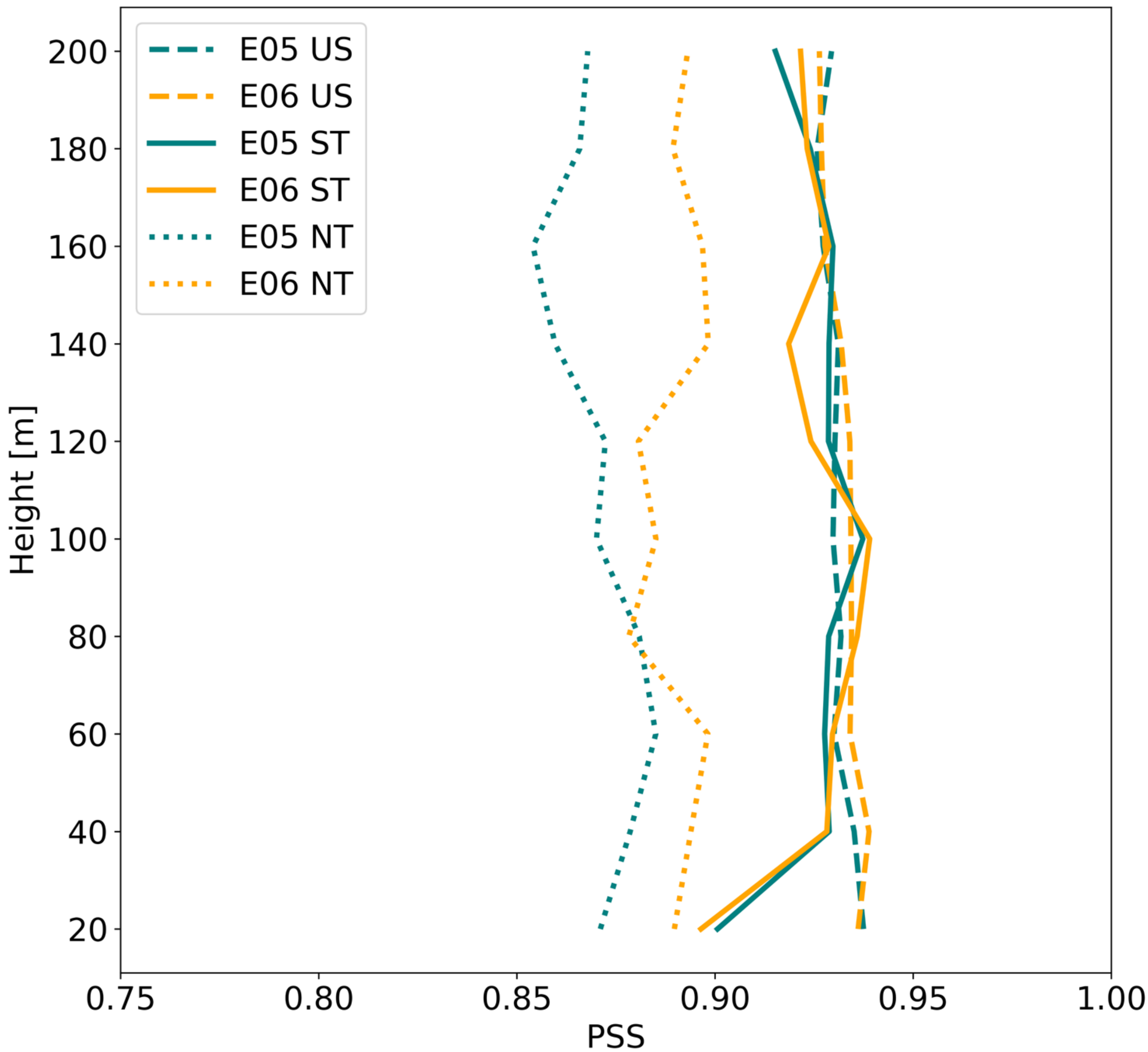


Figure 9. Vertical profiles of the Perkins Skill Score by stratification. at the E05 (teal) and E06 (orange) lidars subset by stratification (US = unstable, ST = stable, NT = neutral).

Profiles of PSS between NWF and lidar observations of wind speed vary by location and stratification. Performance is generally best in unstable conditions at both E05 and E06 lidar locations with a mean value of 0.93. Performance is next best in stable conditions, starting around 0.90 at the surface and increasing to 0.93 at 120 m at E05. At E06 in stable conditions, PSS reaches a maximum value of 0.93 at 100 m. Neutral conditions exhibit worse PSS and larger spread by location. AT E05, PSS minimizes at 0.85 at 160 m and maximizes around 0.88 at 60 m. At E06, PSS scores minimizes at 0.87 at 80 m and maximizes at 0.89 at 140 m.

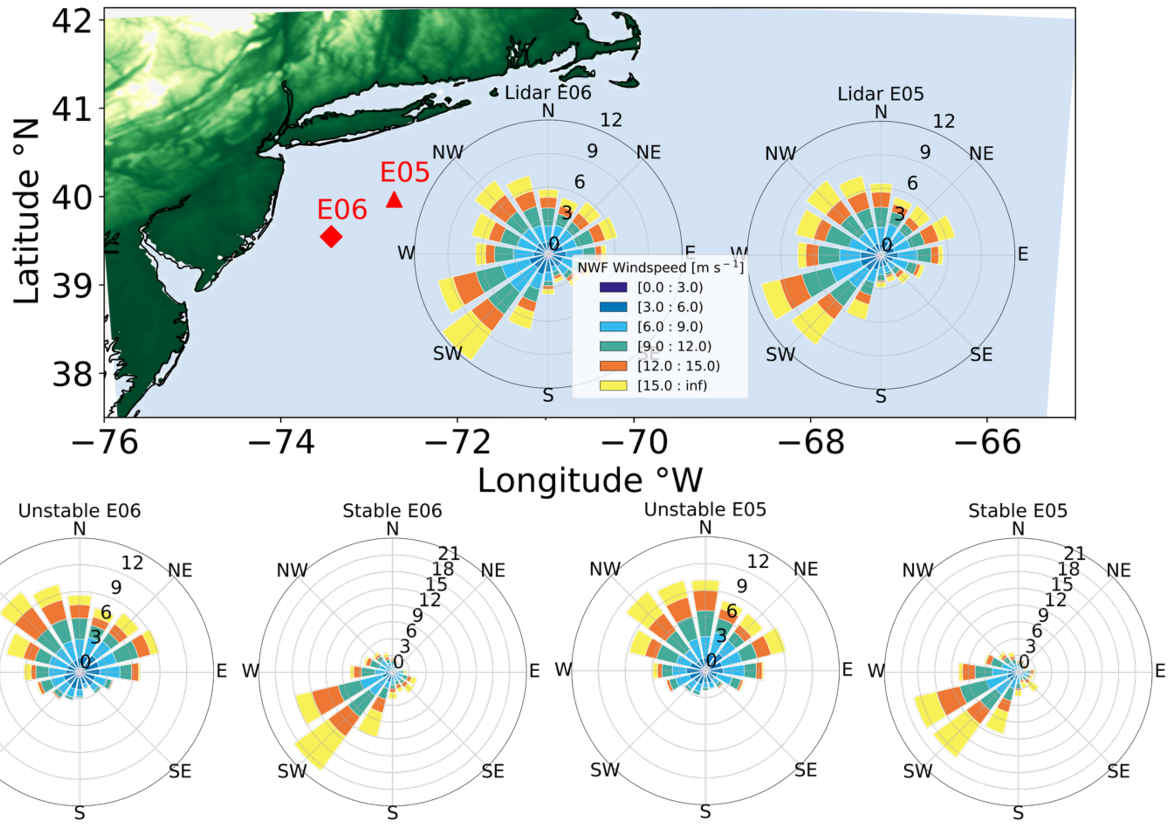
I wonder how the EMD, which is expressed in meters per second, depends on the wind speed itself. Do distributions with higher wind speed also have higher EMD even though the overlap is similar? Does this influence the comparison between stable and unstable that do have different wind speeds?

Distributions with faster wind speeds will not have a higher EMD value. For instance, adding 100 m s^{-1} to every WRF and lidar value, (i.e., keeping the distributions the same but shifting them towards faster values) and recomputing the EMD results in the same EMD value. So, for instance, faster wind speeds in stable stratification will not cause the EMD to be larger, because the same amount of "work" would be required to move modeled data points onto the observed data points.

6. In your analysis about the different stability classes, you mix the effect of the wake with the fact that the wind rose is likely rather different during the different stability classes. I would appreciate a clearer distinction between the effects caused by variations in wake behavior and those resulting from different wind patterns. To analyse this, you have to include wind roses for both unstable and stable conditions. I expect substantial differences, which could be attributed to seasonal variations. To gain deeper insights, stratifying the data according to wind directions might be a valuable addition.

Thank you for the suggestion. We now assess the wake propagation distance to the southeast of the RIMA block in unstable conditions to account for the predominant northwesterly winds.

We have modified Figure 3 to show the relationship between stability and wind direction:



We have modified the methods section to describe the changes on new lines 483-488: "Because wakes typically propagate to the northeast during stable conditions (Figure 3), we calculate the propagation distance of wakes along a line extending northeast of the RIMA block (Figure 1) and report the distance along the line where wake wind speeds reach a threshold. In unstable conditions the prevailing wind direction is northwesterly (Figure 3), so we assess the wake propagation distance to the southeast instead. The threshold of -0.5 m s^{-1} is chosen following Golbazi et al., (2022); Rybchuk et al., (2022)."

We now report wake propagation distances of 3.7 km (TKE_100) and 5.9 km (TKE_0) in unstable conditions on new lines 652 and 654.

7. I find the Turbulent Kinetic Energy (TKE) coefficient sensitivity study interesting and informative. However, I would argue that it does not fully represent uncertainty quantification. There are numerous other uncertain parameters, particularly in wind farm parameterization, which require a more comprehensive approach for uncertainty quantification. I suggest to remove the term uncertainty quantification in abstract and conclusions. You can add a discussion of how this would be a first step towards uncertainty quantification in a discussion.

We agree that a more extensive uncertainty quantification could be considered, but the TKE coefficient receives considerable attention in the literature and therefore addressing its bounding values (0% and 100%) provides a range of variation in wakes due to this parameter and how it affects boundary-layer dynamics and therefore more uncertainty quantification than has yet been addressed. The reviewer's suggestion of emphasizing the "first step" towards a more complete uncertainty quantification is a good one, so we incorporate this language into the abstract and conclusions. New text in the abstract is as follows:

"We also provide a first step towards uncertainty quantification by testing the amount of added turbulence kinetic energy (TKE) by 0% and 100%. We provide a sensitivity analysis by additionally comparing 25% and 50% for a short case-study period."

New text in the first sentence of the conclusions (line 831) is as follows:

"This modeling study assesses the variability of wake effects across the mid-Atlantic OCS based on yearlong simulations, including a first step towards uncertainty quantification and approaches for distinguishing internal and external wake effects."

8. Appendix A, in my view, is very interesting but not yet mature enough for inclusion in the scientific paper. While it is interesting to study the effect of Turbulent Kinetic Energy (TKE) on surface heat and moisture flux, as well as planetary boundary layer heights, it is a different topic that would benefit from additional model evaluation, particularly concerning those variables.

We appreciate that there are other parameters that must be explored for an in-depth sensitivity analysis. The findings shown in Appendix A can provide context for additional questions and research into the sensitivity by TKE amount which has been requested by previous reviewers. This case-study analysis can lay the groundwork for future development. The placement of this material in an appendix rather than the main paper emphasizes that it is supporting material rather than full mature analysis, and so we choose not to remove the section.

9. It would be helpful to provide references or an explanation of the procedure for scaling the 12 MW turbine to the 15 MW turbine.

This scaling was performed by Beiter et al. (2020). We have made this point clearer and have added a brief description and reference on new lines 142-146:
"For our simulations, we parameterize 12 MW turbines *which are* scaled *by Beiter et al., (2020)* from a 15 MW reference turbine with a 138 m hub height and 215 m rotor diameter. *The power and thrust coefficient curves were held constant from the 15 MW machine. The rotor diameter was scaled to maintain a specific power of 332 W m^{-2}, which is the same as the reference 15-MW turbine. Then, the hub height was determined such that a 30-m gap was maintained between the lower bound of the rotor tip and the sea surface.*"

10. I wonder how the wind direction statistics have been conducted, especially given that wind directions range from 0 to 360 degrees. A brief explanation or clarification in this regard would enhance the reader's understanding.

We have added the following text to clarify wind direction statistics on new lines 251-263:

The circularity of wind direction must be accounted for in statistical calculations. For example, computing the average between 359° and 1°, using a typical arithmetic mean, would result in 180°. However, the mean wind direction between those two values should be 360°. The SciPy (Virtanen et al. 2020) and Astropy (Price-Whelan et al. 2022) Python packages offer convenient functions which allow the user to calculate statistics for a circular variable by passing in the lower and upper bounds, in this case 0° and 360°, respectively. We calculate the mean and standard deviation of wind direction using the SciPy "circmean" and "circstd" functions, respectively, and the correlation coefficient using the Astropy "circcorrcoef" function. The cRMSE for wind direction is then calculated using Eq. 9:

$$cRMSE = \sqrt{circmean\left(180° - \left|\left|\left(D_{WRF_i} - \overline{D_{WRF}}\right) - \left(D_{lidar_i} - \overline{D_{lidar}}\right)\right| - 180°\right|\right)^2} \qquad (9)$$

where D is wind direction and $\overline{D}$ is the circular mean of wind direction. Bias is calculated similarly to Eq. 6, except that differences between NWF and lidar values that are less than −180° have 360° added and differences greater than 180° have 360° subtracted (Eq. 10):

$$x = \begin{cases} x + 360° & \text{for } x < -180° \\ x - 360° & \text{for } x > 180° \end{cases} \qquad (10)$$

where x is the $(D_{WRF_i} - D_{lidar_i})$ difference.

11. Regarding Figure 16 and the related description, I am unsure of the significance of the difference between TKE=%100 and TKE=0%. This difference could potentially be attributed to variations in the timing or positioning of weather systems unrelated to the parameterisation. The explanation of Kelvin-Helmholtz instabilities appears speculative and requires a more detailed analysis. I recommend the removal of this analysis from the paper unless more substantial evidence and analysis can be provided.

While we appreciate the concern, the timing of the gust front is dictated by the initial and boundary conditions (which are the same for both simulations).

12. P20: the maximum average wake wind speed deficit: it is unclear what this refers to: is it the maximum deficit in space? Is this within the wind farm or outside? And why do you divide by the average hub-height wind speed to get a percentage rather than obtain the maximum percentual average wind speed deficit? This is not necessarily the same. Would be good to add some clarifications.

Yes, the maximum average wake wind speed deficit is the maximum deficit in space for the mean wakes shown in the figures. The maximum average wake occurs inside the wind farms. Our goal is to show the relationship between the maximum deficit in space and the typical wind speed, as opposed to how large the percentual reduction is when averaged over every timestamp.

We have added clarification on new lines 612-613:
"Here, we categorize wakes by the maximum wind speed deficit *in space*, the spatial extent, and the downwind propagation distance."

We have also added clarification on new lines 614-615: “The maximum average wake wind speed deficit *occurs within the wind plant areas* and intensifies from -1.5 m s^{-1} to -2.8 m s^{-1}, moving from unstable to stable conditions for TKE_100 (Figure 11a,c).”

Referee 2

Referee comments appear in black and author responses appear in blue.

The authors have done an excellent job at addressing my concerns. I only have a few minor requests.

We thank the reviewer for reading the modified article thoroughly and providing suggestions to improve this work.

1. The 25% TKE correction value is the default in WRF. It does not matter that only one study to date (Archer et al. 2020) has supported it. What matters is that *everybody* is using that value unless they have extensive knowledge of the issues associated with it and change it manually. I disagree that "The 0% added TKE is more similar to the impact in the Volker et al. parameterization which has been used in several studies." Volker et al. used a totally different approach for TKE and therefore I disagree that they should be cited in support of 0% TKE.

   We do not wish to claim that 0% added TKE mirrors the approach used by Volker et al., but that Volker et al. rely on wind speed shear to induce turbulence, which shares similarity in concept to the 0% added TKE approach. Regardless, our reference to Volker was only in response to the reviewer and is not used in the manuscript.

   Other studies have used 0% TKE, but only as sensitivity. One of the first papers by Fitch et al. demonstrated that adding 0% TKE was indeed inaccurate for example. We all know that some TKE is indeed added by the turbines and therefore 0% is not a representative value. 100% is the old default and there is value in using it. But not having anything in between is not good. I wish that the authors had done some runs at 25%, but it's too late for that.

   We agree that performing a more in-depth assessment of the variation by TKE amount would be more thorough, however our computational resources limited us to two simulations only. Fortunately, we include a short-term run using 25% and 50% added TKE in the Appendix, which the readers may use to make comparisons and jump start future research.

   But the sentence in the abstract: "We also vary the amount of added turbulence kinetic energy (TKE) between 0% and 100% to provide some uncertainty quantification" is untrue. You did not "vary" the added TKE between the two extremes, you only used the two extremes. Thus the sentence should be rephrased as "To provide some uncertainty quantification, we tested two values of added TKE: 0% and 100%."

   We agree that no TKE values between 0% and 100% were used for the main findings of this study and have modified the sentence according to the reviewer's suggestion on lines 18-

19: “We also provide a first step towards uncertainty quantification by testing the amount of added turbulence kinetic energy (TKE) by 0% and 100%.”

2. I suggest a small change in the order of the equations for the losses, to improve readability (shorter acronyms) and be more consistent (no mix of Loss and LOSS):

$$Loss_{tot} = 100\% - \left(\frac{P_{LA,CA}}{P_{NWF}}\right) \times 100\% \quad (9)$$

$$Loss_{int} = 100\% - \left(\frac{P_{VW}}{P_{NWF}}\right) \times 100\% \quad (10)$$

$$Loss_{ext} = 100\% - \left(\frac{P_{LA,CA}}{P_{VW}}\right) \times 100\% \quad (11)$$

$$Loss_{ext} = Loss_{tot} - Loss_{int} \quad (12)$$

We thank the reviewer for their attention to detail regarding “LOSS” vs. “Loss” and agree that keeping external loss equations next to each other improves readability. We have incorporated the reviewer’s suggestions, in new Eqs. 13-16, having replaced the “VW” acronym with “ONE”, to easily recognize that one wind plant by itself is used, as can be seen throughout the article:

$$Loss_{tot} = 100\% - \left(\frac{P_{LA,CA}}{P_{NWF}}\right) \times 100\% \quad (13)$$

$$Loss_{int} = 100\% - \left(\frac{P_{ONE}}{P_{NWF}}\right) \times 100\% \quad (14)$$

$$Loss_{ext} = 100\% - \left(\frac{P_{LA,CA}}{P_{ONE}}\right) \times 100\% \quad (15)$$

$$Loss_{ext} = Loss_{tot} - Loss_{int} \quad (16)$$

3. I disagree that L=1000 m is a good choice here. The authors provide one reference only for it, Munoz-Esparza et al. (2012), which I am not familiar with. I provided 5 or 6 references for shorter values and recommended 500 m, which is in the ballpark of all of them. Dismissing them because not all of them are offshore is not convincing; not to mention that at least one, Archer et al. (2016), was offshore, more recent than Munoz- Esparza et al. (2012), and obtained from measurements in the Nantucket Sound, which is in the area of interest here. Lastly, using L=500 m the authors obtained a more typical frequency of neutral cases (11.2% versus the previous low value of 4.5%). As such, I have to insist that the calculations be modified using L=500 m.

We thank the reviewer for their suggestion, as our recalculated validation metrics using L=500 m show improvement across most statistics. As such, we have redone all calculations that rely on stratification using the L=500 m threshold.

References

Beiter, P., W. Musial, P. Duffy, A. Cooperman, M. Shields, D. Heimiller, and M. Optis, 2020: The Cost of Floating Offshore Wind Energy in California Between 2019 and 2032. *Renewable Energy*, 113, https://doi.org/10.2172/1710181.

Bodini, N., and Coauthors, 2023: The 2023 National Offshore Wind data set (NOW-23). *Earth System Science Data Discussions*, 1–57, https://doi.org/10.5194/essd-2023-490.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. *Journal of Climate*, **20**, 4356–4376, https://doi.org/10.1175/JCLI4253.1.

Price-Whelan, A. M., and Coauthors, 2022: The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package*. *ApJ*, **935**, 167, https://doi.org/10.3847/1538-4357/ac7c74.

Pronk, V., N. Bodini, M. Optis, J. K. Lundquist, P. Moriarty, C. Draxl, A. Purkayastha, and E. Young, 2022: Can reanalysis products outperform mesoscale numerical weather prediction models in modeling the wind resource in simple terrain? *Wind Energ. Sci.*, **7**, 487–504, https://doi.org/10.5194/wes-7-487-2022.

Redfern, S., M. Optis, G. Xia, and C. Draxl, 2023: Offshore wind energy forecasting sensitivity to sea surface temperature input in the Mid-Atlantic. *Wind Energy Science*, **8**, 1–23, https://doi.org/10.5194/wes-8-1-2023.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Xia, G., C. Draxl, M. Optis, and S. Redfern, 2022: Detecting and characterizing simulated sea breezes over the US northeastern coast with implications for offshore wind energy. *Wind Energy Science*, **7**, 815–829, https://doi.org/10.5194/wes-7-815-2022.