In this document, the reviewer's comments are in black, the authors' responses are in red.

We thank the reviewer for their thoughtful comments, which gave us an opportunity to revisit our analysis.

The manuscript by Bodini et al. on "Assessing Boundary Condition and Parametric Uncertainty in Numerical-Weather-Prediction-Modelled, Long-Term Offshore Wind speed Through Machine Learning and Analog Ensemble" focusses on comparing two different methods on how to extrapolate the ensemble uncertainty of a short (one year time) series to a long time series for the wind conditions along the coast of California.
The study is generally well written, figures are selected meaningfully and readable. However, I have a number of minor comments to be taking into account before I can recommend publication in Wind Energy Science.

Minor Comments:

- Line 28: Northern Europe is typically referred to as Scandinavia. I guess you rather mean central Europe here?
We have rephrased this as "In some areas of the world, such as the North Sea in Europe,"

- Line 43: "a continuous, in space and time" → I suggest: an in space and time continuous
Changed.

- Line 76: "prohibitive, and innovative and more computationally" → remove the "and" before innovative
We hare rephrased this.

- Line 103: "The 15 WRF ensemble members" → I strongly recommend creating a table here with one row per one of the 16 members and then the settings in the columns. It is hard to read and understand here.
We had a typo in the sentence. Year 2017 (and not the 15 ensemble members, as previously stated in the draft) was selected because of strong data coverage from the network of buoy and coastal radar observations used for model validation. We think that now the section is clear: all 16 WRF ensembles share the common WRF attributes in Table 1; the list at page 5 shows the various combination of additional WRF parameters that led to the creation of the 16 ensemble members.

- Line 131: "mean hourly wind speed" → How did you compile the mean hourly wind speed? How many timesteps have you written out from the WRF runs to compute the average?
We have added the following: "(calculated from 5-minute WRF raw output)".

Line 147: "the WRF across-ensemble standard deviation" → what is the across-ensemble standard deviation. Is this the standard deviation from the ensemble members?
The WRF across-ensemble standard deviation is explicitly defined in equation 1.

Line 156-157: Can you explain why you used both the standard deviation from preceding 2 and 6 hours and how these time intervals were selected?

We selected these time intervals to try to capture variability of atmospheric conditions over different time scales which based on our experience could have an impact on hub-height wind speed uncertainty. As stated in our answer to the next comment, we have acknowledged in the paper that the choice of input features in our analysis was not exhaustive, and that "Testing additional input features to the algorithms could also help further improving the accuracy of the proposed extrapolation."

Line 161: "However, we found that including all the features" → can you explain how you found this? And did you try further features?
We compared the error metrics obtained (on a limited number of test sites) when using different sub-sets of input features, and found that the best results were obtained with all inputs together. We did not try additional features, but we agree that this might be interesting future work, and we have added the following sentence to the Conclusions: "Testing additional input features to the algorithms could also help further improving the accuracy of the proposed extrapolation."

Table 2: It is quite difficult to read column two. Maybe it helps to add a small empty row after each row?
Done.

Table 4: I think you should at least try to explain why there is no weight on the inverse Obukhov length but a strong weight on shear.
As suggested by the other reviewer, in stable conditions the depth of the Prandtl layer is much shallower than turbine hub height so that the near-surface Obukhov length is likely irrelevant for determining uncertainty in the 100 m wind speed. We have now added some comments on this in the paper.

Line 230 and Line 303-304: "Applying the bias correction proposed in" → So, if this would likely reduced the AnEn bias, why didn't you do it? This should at least be explained.
The negative bias is likely caused by two reasons, the limited historical repository to search and the fact we take an *average* of the 16 ensemble members. When only having a limited historical repository, there is a balance between the prediction accuracy and the systematic bias. During our grid search analysis for the best number of ensemble members to generate, we observed that AnEn archives better bias (closer to zero) with fewer members but the prediction accuracy, e.g., RMSE, would drop. To the focus of this manuscript, prediction accuracy is the most important metric, and therefore, we set RMSE optimization to be the most important metric. Additional bias correction processes might be able to help AnEn prediction, but we would like to keep a fair comparison between ML and AnEn. In fact, the postprocessing technique can be computationally expensive when applied to a large domain. Therefore, extensive correction on AnEn has not been explored.
We have added information on this in the following paragraph, where we have also added reference to another technique that can be explored in future work to reduce the AnEn bias:
"The bias from the AnEn approach is likely because of the reduced length of the search period (1 year), which might be too limited for identifying a significant number (16) of analogs. This setup constrains the AnEn ability to account for rare events (e.g., particularly high-wind-speed cases) when looking for similar atmospheric conditions in such a short repository. Also, when searching for the optimal number of analogs to use, there is always a trade-off between the prediction accuracy (e.g., the RMSE) and the prediction bias. For our analysis, the main goal was to maximize

the prediction accuracy, in alignment with the ML approach, and therefore we set the RMSE as the optimization metric. During our grid search analysis to determine the optimal number of analog members, we observed that AnEn archives better bias with fewer members, which would however worsen the prediction accuracy (RMSE). Applying the bias correction proposed in Alessandrini et al. (2019), using a machine-learning similarity for analog definition (Hu et al., 2021b), or adopting a quantile mapping that uses quantiles of the analog ensemble instead of its mean (Sidel et al., 2020) would help reducing the AnEn bias, at the potential expense of computational costs."

Figures 6 and 8: There should be a space between caption text and unit.
Thank you for catching this, we have fixed this.

Line 279-280: "For each atmospheric stability class, the values shown are the" → Maybe an active sentence is better here, e.g., for each atmospheric stability class, we show....
Changed.

Line 293 and the whole Conclusions section: I suggest replacing Weather Research and Forecasting or the abbreviation WRF by "mesoscale" everywhere in the conclusion, because the results and conclusions should in principle be valid to any mesoscale ensemble dataset, shouldn't they?
We agree with the reviewer and have changed this as suggested.

Line 319: "deployed in the California OCS very recently" → I guess you mean this buoy? https://a2e.energy.gov/data/buoy/lidar.z06.00 Maybe it's worth referencing the dataset (including the doi) here.
We have added a reference as suggested.

References: The references are vastly incomplete. Every journal paper should have a doi and every technical report a link or other accessibility information.
We have updated the references whenever possible.