Overall Summary and Impressions for the Author

The manuscript investigates offshore hub-height wind predictions for the U.S. Pacific Outer Continental Shelf, specifically focusing on the Humboldt and Morro Bay Wind Energy Areas (WEAs). The study compares the performance of machine learning (ML) models, including Random Forest (RF), Gaussian Process Regression (GPR), and Long Short-Term Memory (LSTM) neural networks, with the traditional stability-corrected logarithmic law (S-C log law) for predicting hub-height wind speeds at 100 meters above mean sea level (AMSL). Input features for the ML approaches are derived from sea surface measurements and supervised by lidar measurements of winds aloft collected from DOE-deployed buoys. The models are trained on data from one location and tested on data from the other, a method used in other wind speed extrapolation ML models (Bodini & Optis 2020) to minimize bias and simulate realistic applications. The study reveals that the S-C log law and LSTM models have the best prediction accuracy. ML models advantage over the physical algorithm is the capability to generate additional insights, such as predicting turbulence intensity (TI) and providing confidence intervals in the case of the GPR model. Additionally, challenges like computational expense in algorithms and the influence of atmospheric stability conditions and location on model prediction accuracy are addressed.

The study introduces a comparison of physical modeling and ML-based wind prediction aloft, which appears valuable for advancing offshore wind energy applications. Overall, I think providing additional clarity and explanation of the author's methodology and results would dramatically improve the strength of this manuscript. I would like to see the results section strengthened to showcase the benefits of the ML models. While I think my comments are minor overall, I would recommend major revisions to offer the authors more time to address comments.

Major Comments

Line 95 & overall comment- the idea of training and testing the models at different data locations is intriguing, but if data limitation was a big limitation in the study, why not create one more robust model trained and tested on data from both sites to ensure universality using methods such as k-fold cross validation (suitable for small datasets) or time series cross validation (good for including neighbor information)? The authors may want to try out this methodology and compare to their current models. At minimum, can the authors explain why this methodology wasn't explored? Why also did the authors create 6 models

from monthly data instead of one model that includes data from each month? I think future analysis could benefit from incorporating other methodologies.

Line 130 & general comment to address- authors discuss filling nan values with mean hourly values. Can the authors discuss what biases they might have introduced to the data? Another question. LSTM and GPR models use neighbor information, where this nan filling may have been required, but RF doesn't. Did all 3 models use the same input dataset or were the filled values excluded from the RF dataset? Also please explicitly state the size of the datasets. This comment also feeds into a later experiment described in this manuscript. The authors trained at one site and tested at another, and vice versa. However, the Humbolt dataset was much smaller in size. For comparing these models, were they each trained on the same size dataset?

Line 134- What values did you clip the data to? And did it differ between the two sites?

Line 166 and general comment- In my opinion, the section on the atmospheric stability calculations came out of no where. It wasn't mentioned in the abstract or introduction. This part of your analysis needs to be explicitly stated to help guide your reader.

Table 4 & general comment- What values did your tuned algorithm use? You tuned between the range of values given, but what did you use for your final model? General comment, in the manuscript, the author mentions many times that further tuning could improve the models, but this table along with this repeated comment gives the impression the author didn't put significant time into tuning the model. Presumably the parameters underwent rigorous tuning and the best model was used for analysis.

Line 245- Is the same scaling used for the GPR and RF? If not, why?

Figure 7- could you include a fourth subplot showing the input data? At the very least, surface wind speed and air-sea temperature difference, as they were indicated to be the two most important featured variables. Could the authors comment on this case why all three models underpredict wind speed compared to the observations and describe the atmospheric conditions during this period?

Line 325- Can you provide some statistics of performance based on different conditions? For example, accuracy day versus night, stable versus unstable. This would make your results more robust rather than looking at a timeseries representing one day.

Figure 10 and line 342 and 352- I would be interested to see figure 10 replicated for the same location train-test models, perhaps in a supplementary section.

General comment 3.1.3 Overall results show that the LSTM does on par with the S-C log law. However, I'm curious if there are specific cases in which the ML models perform

better. A demonstration of this phenomena would strengthen the paper's conclusion. Perhaps the authors could include a sample time series with the surface data showing the onset of a cold or warm front if it shows that the ML models are more adaptable to forecasting the changes aloft. I recognize this analysis may be beyond the scope of this paper.

Line 377- Have the authors considered a neural network with input channel dropout layers to improve the model's elasticity with missing data?

Line 392- Can you use the lidar's data to confirm sheer conditions?

General comment for analysis based on figure 2- Could you compare the distributions of windspeed from the ML models and physical algorithms?

General comment for analysis based on section 3.1.5- Out of curiosity, could you compute error metrics but separate it by wind speed? Maybe the ML models excel under different wind conditions.

Minor Comments

Line 19- Paraphrasing your abstract, you say ML techniques… can be used to predict other wind parameters (plural), but in your paper the only one you evaluate is turbulence intensity. I would be explicit here to not overrepresent your results.

Line 90- provide citation

Line 198- cite the modified log law

Section 2.5- for clarity can you state the lidar buoy data is used as the supervised output dataset for training the model and the S-C log law result is for comparison?

Line 315- describe EMD for general audience

Line 357- Regarding improvement of LSTM over S-C log law, can the author add numbers here?

Line 367- state accuracy

Line 431- I had trouble understanding this sentence. Were the surface variables significant or did feature importance show they were all insignificant?

Line 486- mention this limitation earlier when describing the dataset