Response to Reviewer 2

This paper introduces a decision tree-based application of the measure-correlate-predict methodology, which is commonly employed in the wind industry, to estimate wind gusts. The manuscript is very well written and provides interesting historical context. The machine learning approaches outperform ERA5 in wind gust representation at three observational locations in West Texas and predictor variables are ranked in order of importance to the algorithms.

We thank the reviewer for the succinct summary and positive remarks.

The manuscript, while quite interesting, would benefit from some discussion about the potential applicability of these gust estimation techniques to the wind energy audience of this journal. To elaborate, the tests performed in this work are at 10 m above ground level, while typical turbine hub heights are 80+ m above ground level. Do the authors have any speculation or insight into how the performance of their models and the importance scores of the various parameters might change for a higher height?

We have elaborated on this issue in Section 8 of the revised manuscript.

Additionally, it would be helpful to understand the ultimate goal of this research to support the wind community. Is the aim to improve gust estimates in wind farm forecasts? Or to provide long-term assessment of the frequency and magnitude of gusts at a proposed wind farm? Lines 198-201 hint at the latter, however, an explicit goal statement would be advantageous to the text.

In the abstract, we clearly mentioned that our goal is to generate long-term, site-specific wind gusts data via machine learning. Towards the end of Section 3, we also mentioned that we are proposing a measure-correlate-predict (MCP) approach for wind gusts. See also the first sentence in Conclusions.

In addition to site assessments, our proposed INTRIGUE approach can be used in wind gust forecasting. Instead of a reanalysis dataset, predicted meteorological fields from a numerical weather prediction model can be used as input features for the ML models. We added this information in the revised manuscript.

My one concern with the analysis is the method selected for comparison of time-series of varying temporal resolution (Lines 182-187). Taking the maximum gust from the 5-minute observations within each hour and assigning it to the top of the hour for evaluation of hourly models seems an unfair comparison, which is particularly noticeable in Figure 5. Was there a reason you did not choose the observed wind gust closest to the top of the hour for your comparisons?

The reviewer raises an important issue. Unfortunately, there is no fool-proof approach in the literature for comparing high-frequency point observations and numerically modeled data that are spatially filtered. Our strategy had two justifications.

First, the proposed decision tree-based approaches utilize the variable W_{p10}^{m} (called $10fg$ in ERA5) as one of the input features. This particular variable represents mean wind gusts during the past hour and, thus, is expected to be correlated with the observed maximum gusts during the same period.

Second, the observed wind gusts are spatiotemporally intermittent. Therefore, we were concerned that by simple temporal downsampling, we would lose a significant number of extreme gust events. Averaging the observed 5-min gusts to the hourly averaged value was not considered a viable alternative, as this approach would have drastically reduced the amplitude of extreme wind gusts. Thus, we opted for taking the maximum value over the past hour. We would like to point out that each grid point of ERA5 effectively represents a spatial coverage of approximately 32 km x 32 km; hence, wind gusts from the ERA5 reanalysis are expected to be much less intermittent than observations.

Other comments: Line 102: “Quasi-universal” is a bit of a strong statement given the small sample size.

We changed it to ‘not site-specific’.

Line 177: The authors should consider elaborating on why ERA5 was selected, including citing wind studies that employ it. Additionally, I think a literature review discussion on the accuracy of some of the ERA5 variables employed as predictors in the analysis, in particular the ERA5 instantaneous wind gust, friction velocity, and boundary layer height.

Soon after its introduction, the ERA5 dataset became the preferred reanalysis dataset in the wind power meteorology community. The following papers (and many others) discuss its superior accuracy, lower uncertainty, and higher reliability compared to other global reanalysis datasets.

1. Olauson, J. (2018). ERA5: The new champion of wind power modelling?. Renewable Energy, 126, 322-331.
2. Ramon, J., Lledó, L., Torralba, V., Soret, A., & Doblas-Reyes, F. J. (2019). What global reanalysis best represents near-surface winds? Quarterly Journal of the Royal Meteorological Society, 145, 3236-3251.
3. Gualtieri, G. (2022). Analysing the uncertainties of reanalysis data used for wind resource assessment: A critical review. Renewable and Sustainable Energy Reviews, 167, 112741.

We have included these references in the revised manuscript. It is beyond the scope of this paper to discuss the accuracy of the ERA5 variables.

Line 180: What is the distance between each observation station and its nearest ERA5 point?

The distances between the REESE, MACY, FLUVANNA stations and their corresponding ERA5 grid points are 14 km, 9 km, and 12 km, respectively. We have added this information in the revised manuscript.

Section 6.4: I encourage adding bias to the list of performance metrics, as it would be valuable for the wind community to understand whether the evaluated models tend to overestimate or underestimate observed gusts.

We have added bias as an additional performance metrics.

Line 268/Tables 3-5 (and 7): "From Tables 3-5, it is clear…" would be clearer if these were figures instead of tables. For example, each one could be a line plot with subplots a, b, c for the different stations.

Following the suggestions of both reviewers, we have replaced the tables with bar diagrams in the revised manuscript.

Line 278: "In Tables 3– 5, all the scores of the ML models are averaged over ten years. Thus, the inter-annual variability of all these models is much lower in comparison to the ERA5 baseline." These sentences are confusing. Is the inter-annual variability is lower because all of the scores are averaged or because of the model performance?

We agree that the original sentences were a bit confusing. We have rewritten them in the revised manuscript as follows:

> "In Figures 5–7, all the scores of the ML models are averaged over ten years. Due to averaging, the perceived inter-annual variability of all these models is much lower in comparison to the ERA5 baseline. For example, in the context of the XGB model, the R2 score at MACY has a narrow range of 0.68–0.70. In order to investigate the year-to-year variability and performance of an ML model, we report the annual R2 scores at the MACY station in Table 3."

Section 7.2: This section is quite interesting, given the challenges of observational coverage. I encourage investigation of geographic range of applicability of the discussed techniques for a future paper.

We do intend to expand our work in this direction.