Responses to reviewer comments

Reviewer comments are in black. Responses are in purple. The authors thank both reviewers for their detailed reviews and suggestions for how to improve this work. Major changes have been made including the following:

1) Improved explanation of the approach and models used.
2) Developed and included comparison with a cosine power law function for the power as a function of wind vane.
3) Included comparison of ML predictions with a simpler linear regression model.
4) Updated the ML approach to use a gradient boosting regression model which is a similar ensemble of trees while also enabling quantile regression. Thus, all of the results have been updated with the new model.
5) Added prediction uncertainty estimates using quantile regression.
6) Sensitivity analysis has been added for the consensus algorithms.
7) Correction of the citation format throughout the paper.
8) Revised the discussion and conclusion to reflect the updated results.
9) Included several additional references (note these are not shown in the diff of the manuscript due to how this was created). The added references are: Astolfi (2024), Astolfi (2023), Liew (2020), Fezai (2020), Zhang (2018), Friedman (2001), Chen (2016), Koenker (2017), Li (2014), Bergmeir (2019), Virtanen (2020), Breiman (2001).

Reviewer 1:

The manuscript develops a machine learning model to predict the wind speed, wind direction, and power production yawed turbines would experience if they were unyawed, with has as inputs only SCADA data. To support this, it also develops a way to estimate these quantities based on data from neighboring turbines. This work is useful, as it addresses a real problem that will become more important as active wake steering becomes more widespread. Furthermore, the use of real SCADA data demonstrates that the method could be implemented in real-world scenarios. However, the developed model itself is not convincing, and the manuscript does not explain it well. A full overview of the issues I have is given below.

Overall, the manuscript needs major changes before being considered for publication. I recommend adding a simpler model for comparison and a major rewrite of some sections, as discussed below. I would be happy to review a revised version with these changes.

Thank you for your detailed review and recommendations. We have worked to improve the explanation of our modeling approach. As suggested simpler models have been included for comparison. One of the drawbacks to simpler models we now discuss in the revised manuscript is that they don't have the ability to include additional features (e.g. wind speed or power, turbulence intensity). Therefore while some comparisons can be drawn and the relative accuracy of the models compared, these comparisons are not exactly apples to apples.

Major comments:

1. The authors use a machine learning algorithm as their model for predicting the unyawed QoIs based on the yawed measurements. They argue that this is necessary to enable non-linear fitting of the NTFs (line 174). However, later results such as figure 9 seem to give quasi-linear predictions for the majority of the inputs.

   It therefore seems to me that a simple linear regression between the measured and predicted wind vane and speed could perform well. Furthermore, literature already contains models relating yawed and unyawed power output as a function of turbine angle, such as a cosine power law. Compared to these simpler approaches, the full ML algorithm developed by the authors seems needlessly complex.

   Including such a simple approach and comparing it against the ML model would greatly enhance the value of this manuscript. If not, the authors should argue more strongly and convincingly why a fully non-linear model is needed, building on earlier literature.

Analysis using simpler linear regression and cosine-power-law modeling has been added for comparison and discussion.

2. The bagged tree regression model used by the authors is not explained in the text, outside of a few sentences giving a conceptual description (lines 175-180). This is not sufficient, and makes it impossible to understand the work without consulting external sources. Please add a description of the algorithm, along with the core equations if possible.

Thank you for your comment. We have revised our analysis to utilize a gradient boosting regressor (GBR) instead of the bagged tree regressor in the original version. GBR is still an ensemble of trees similar to the bagged tree regression that was previously used but has some advantages because it facilitates a unified uncertainty quantification approach across the three different types of models we studied (ML, linear regression and cosine-power-law regression). We also included a new paragraph and the core equations outlining the major algorithm components to build the GBR.

3. Overall, the manuscript is poorly written. My main complaints regarding the writing were:
   1. The authors do not differentiate between the \citet and \citep LateX commands, resulting in each citation being a full name citation, without brackets. Normally, this would be a minor comment, but since it happens for every single citation it makes the first two sections of this paper unreadable. The paper cannot be accepted without correcting this.

Apologies for this error in our draft manuscript. It occurred due to a late change in the template used and we did not catch that the references were not appropriately separated from the text. This has now been corrected.

2. The “wind vane” QoI is defined as the “relative wind direction sensor” (line 54-55). It is not immediately clear whether this corresponds to wind direction, turbine yaw angle, or some combination of the two. This is complicated by the authors not being consistent in their usage of the term and reverting back to “wind direction” several times throughout the manuscript (eg. figure 2 caption, line 234, line 258, to name but a few). Please differentiate better between yaw and wind direction.

   Wind vane is the relative wind direction. The true wind vane can be defined as the difference between the wind direction and nacelle yaw position in a global reference frame. The measured wind vane is what an actual mechanical vane or sonic anemometer mounted on the nacelle measures. We have now included a terminology section to clarify our terminology for the reader and worked to ensure consistency throughout the paper.

3. There is no single paragraph clearly stating the goals, methodology, and structure of the paper. The final sentences of sections 1 and 2 give the goal, but only in section 3 is the overall methodology outlined. The manuscript would be greatly improved by a paragraph in section 1 listing how the authors will tackle the problem and which sections of the paper discuss what.

   In an effort to improve clarity, we have expanded the paragraph in section 1 to include both the goals and the methodology.

4. The workflow of this model development is quite complicated and difficult to follow. There are no equations, so it’s hard to see where one model ends and another begins. The manuscript would benefit from some sort of data flow diagram, visualizing what inputs are used by which model and what outputs are produced.

   Excellent suggestion. A flow chart of the approach has now been included to supplement the description.

4. Uncertainty is not considered throughout the manuscript. I suggest the authors incorporate this into their model, as this would greatly enhance its useability for industry applications. If this is not possible, the discussion around model uncertainty should be more in-depth.’

This is an excellent suggestion. The authors modified the analysis to utilize ML models that natively incorporate quantile regression that we utilize to represent the prediction uncertainty.

5. Figures 9-11: Building on the previous comment, these plots would highly benefit from uncertainty estimates around the predictions. Based on the results presented here, I am not convinced that the non-linearity for high wind vane angles in figure 9 is not a statistical artifact, since based on figure 6 there does not seem to be much data available at these angles. Please discuss this further.

   Based on access to new models, these plots have been recreated in a manner to highlight the uncertainty prediction. The discussion in the paper has been updated based on these results.

6. Figure 11: A direct comparison with a cosine power law would be very interesting here. Consider adding that to the paper.

   As suggested, a comparison and corresponding discussion has been added.

7. Line 181: The authors mention that the ML model can select predictors or input variables. However, the only results they show with this seems to correspond to a sensitivity analysis for the different inputs, which many models, including simple linear regression, can do. It's very unclear what the added capability of the developed model is in this context.

   In the revised manuscript, we have taken care to use more precise language around the sensitivity analysis that was performed. As the reviewer has stated, this is not a unique feature of the ML model selected.

8. Line 215: The calculation of the turbulence intensity is not clearly explained. Are the "global average" and the deviation taken over all the turbines for a given timestamp? Or are these values calculated separately for each turbine? In case of the former, the calculated value might be more indicative of general flow non-uniformity throughout the farm, and not just turbulence intensity, which would greatly change the interpretation of some later results. For instance, the discussion on line 222 would be moot, since the high circular differences would then trivially correspond to flow non-uniformity.

   Furthermore, it's not clear whether this is the standard method for estimating turbulence intensity based on SCADA data. If so, please refer to relevant literature.

   The authors have clarified the calculation of turbulence intensity in an updated section 3.1 to include a more complete description of the data including turbulence intensity.

Technical issues and minor comments/suggestions:

1. Were the QoIs considered here (vane, speed, and power) the only available data? Or was this a selection by the authors?

These quantities were selected as three measurements that are critical for evaluating wake steering performance. In addition, these QoIs are potentially influenced by yaw misalignment.

2. Line 188: what is the difference between a predictor feature, a parameter, and a model input? Please be more consistent in the terminology.

As the reviewer states, the terms for predictor features, parameters, and model inputs are not used consistently. In the context of supervised regression, the predictor features are the inputs to the gradient boosted regression model and the linear regression model. We have tightened the terminology for clarity to refer to the model inputs more precisely.

3. Line 196: “base estimators” are never defined. I assume this is related to major comment #2, but a clear description would be appreciated.

Base estimators refer to a single decision tree trained using the available dataset to perform supervised regression. In response to major comment #2, the authors have included a description of the major components of CART based ML models, and provided a clearer description of the base estimator.

4. Figure 2: This figure shows the data before the filtering operations. This should be clearly mentioned in the caption.
   As a suggestion, consider moving this figure to section 3.2, where the filtering is developed.

Figure 2 shows the data before the filtering operations, and for clarity, the authors have changed this to a kernel density plot, rather than a scatter plot and moved this figure to section 3.2, as suggested.

5. Line 212: “stable” conditions has the connotation of stratification effects. Since the authors are discussing non-uniform conditions, consider using “uniform” instead. The same comment applies to line 333, and to “steady” on line 218 (implies transient effects).

Based on the reviewer’s comment, the authors have changed this word selection to reflect that this term refers to spatial consistency.

6. Figure 3: Turbine 456 is not plotted in figure 1. Why is this?

Our analysis focused on the first layer of turbines - a subset of the larger farm. We have now updated the figure to focus on the attention on the first layer of turbines.

7. Line 239: “While difficult to read due to the sheer number of datapoints” using a transparent scatter or a KDE would solve this issue, and both are done later in the manuscript. Please just do that here as well.

   This is a good recommendation. To unify the figure construction, the scatter plot has been updated to a transparent KDE plot.

8. Figure 4: it’s not clear what value this figure has over figure 5.

   This is an excellent point, we have consolidated the figures together.

9. Line 266 / Figure 6: Wind speeds below 5m/s are filtered out. Please just limit your plot axes to this same range.

   This is a good point, we have updated the figure.

10. Line 263: “Inferred by the NTF” does this mean figure 6 shows the dataset the NTF is trained on? Or does it mean that figure 6 shows the output of the NTF? This wording is confusing.

    The figure displays the KDE density of the data on which the NTF is trained. The text describing the figure has been updated to provide clarity.

11. Figure 6: The horizontal line corresponding to the desired 1 ratio is not indicated in several subfigures, which makes the results harder to read.

    This helpful addition has been added to the figure.

12. Line 301-302: the figure is not misleading, as the second column solves the issue you mention here.

    The description of this figure has been updated. What we intended to state is that the model is doing a good job of identifying the center of the data without bias.

13. Figure 8: What is the third column showing? There is no reference to it in the caption or the text.

    The third column shows the theoretical quantile prediction plot. Text has been updated.

14. Figure 8, bottom left: Why is the distribution slanted? There are clear borders, and parallel lines of points within the distribution.

    The turbines are power limited on the high end by the rated power that can not be exceeded by much. On the low end, we filtered out low power data below 200 kW from the data set to avoid the large noise in power ratios at very low power levels. However, the density shows that there are not many data points at these extremes. A discussion of this has been added to the paper.

Reviewer 2:

This manuscript presents a SCADA-based approach for estimating nacelle transfer functions (NTFs) describing the effect of intentional yaw misalignment on measured wind direction, wind speed, and power. The authors first test several consensus-type estimators that use neighboring
turbines to infer yaw-aligned operating conditions and then use the selected estimator to generate reference values for training a machine learning model. The ML model maps measurements from yaw-misaligned turbines to the inferred yaw-aligned quantities.
The topic is relevant, particularly in the context of increasing use of wake steering and the lack of external inflow measurements at many wind farms. The dataset is substantial and the general idea is interesting. However, there are several issues with the methodology that currently limit confidence in the results. Additionally, the lack of benchmarking against existing methods for modeling yaw-misaligned wind turbines makes it difficult to assess the contribution of this work. The manuscript is also difficult to follow due to unclear terminology and insufficient explanation of methods and results. Significant improvements in methodology, benchmarking, and clarity are necessary before the manuscript can be considered for publication. I would be happy to review the manuscript again once these issues have been addressed.

Thank you for your detailed comments. We have worked to correct the deficiencies you have identified in our revised manuscript including improving the explanation of the methodology and including benchmarking against other models.

Major Comments
1. The proposed ML-based NTFs are not compared against any established models for yaw-misaligned wind turbines. There is no comparison to simple cosine law relationships which are the state-of-practice method (Gebraad et al. 2016). Additionally, there is no benchmarking against conventional NTF calibration methods based on met towers or LiDAR. Since the stated goal is to replace met-tower-based calibration, the absence of any validation against a met-tower-derived NTF significantly weakens the contribution. Without such benchmarks, it is difficult to judge whether the proposed model is useful in place of existing methods. These existing methods should be discussed in the introduction and at minimum a cosine model should be used for a baseline comparison.

Unfortunately, Met tower and LiDAR data was not available to the authors for this data set. A Cosine law is potentially appropriate for modeling power as a function of yaw misalignment. Linear relationships can be applied for wind speed and vane angle deviations. Both have now been included in the paper as comparisons.

2. A key assumption in this work is that yaw-aligned turbine quantities can be reconstructed reliably from neighboring turbines at one-minute resolution. Temporal and spatial variability can lead to differences between turbines that are unrelated to yaw misalignment. In turbulent flow, averaging may reduce noise more effectively than aggressive filtering. A sensitivity study to time-averaging the data using various averaging windows (e.g. 2 min, 5 min, 10 min, etc.) could be performed.

It is true that temporal and spatial variation in wind will create challenges for the evaluation of neighboring turbines. Due to temporal variation of the wind direction as well as the wake steering applied, vane angles are also not constant. Extended time averaging will tend to blur the impact of applied vane angles which we are trying to measure. 1-minute resolution is a good compromise. When fitting a model to the resulting data errors should average out as long as the time interval does not introduce a bias. Shorter data time periods should be less likely to introduce such biases.

3. The final NTF model is trained using the full dataset (L194, L204) and as a result the model is not tested on any unseen data. The best practice for ML model development is to use separate datasets for training, validation and testing. Since the model is not tested against a holdout test dataset it is impossible to know if the model has been overfit to the data and therefore how it will perform on unseen data. As a result “testing RMSE” in Table 6 is misleading since the model has been trained on the test data. This is a major oversight in the ML methodology of the work. The model should be trained, tuned, and tested on three separate subsets of the wind farm data and the model should never receive information about the testing dataset before the final evaluation.

It is true that care must be taken to evaluate the performance of machine learning models on test data that they have not seen during training. In data rich environments, it is typical to segment or randomly select a portion of the data to use for testing that the model will never see in training. However, an alternative best practice that is used in our paper is the k-fold approach for evaluation of models and it has some distinct advantages in this case.

There are several challenges that must be addressed when applying machine learning to wind plant data:

1) The data is highly autocorrelated in time (one data point is very similar to the point before or after it).
2) The data is seasonal and weather pattern driven so some combinations of features may only appear in one time portion of the data set.
3) The data is relatively small (from a machine learning perspective we don’t have a lot of data).

Taken together these factors strongly motivate using a k-fold approach to evaluating model performance. To overcome (1) the data is segmented in time rather than randomized which would cause the test data to look nearly identical to the training data. To overcome (2) we could randomize the data but that would cause problems for 1 so instead, we train our model k times where all but one of the folds is training data and the k-th fold is the test data. This way a trained version of the model has to predict each “test” segment in time using only the training data from the other time segments. To address (3) we consider the goal of the approach which is to use the best possible model as our NTF for future wind farm wake steering control, etc. In the case of this paper that is represented for selected conditions by our final NTF plots. The best model for that purpose is one that is trained on all of the available data.

The resulting RMSE reported in table 6 is the residual test RMSE from the combined k-fold applications of the model where it predicts data on a portion of the time series that it has not been trained on. By using the identical hyperparameters for each version of the model trained on each fold, as well as for the final model that trained on all of the data we can ensure that the model does not overfit the training data. Taken together all of the k-fold test segments represent the entire data set and this RMSE provides an upper bound on the RMSE of the model when it is subsequently trained on all of the data. We have now included a new Figure 5 to help explain the k-fold analysis to do the machine learning validation portion of the study.

4. The parameters of the weighted averaging methods (e.g., Gaussian width, number of clustered turbines) appear ad hoc (Table 2). A sensitivity analysis could be helpful to justify these choices.
To select those parameters, we originally did an informal sensitivity study, however we didn't include that in the draft manuscript. We have now formalized this study and included the results showing how these parameters were selected.

5. Figures 5 and 6 contain many subfigures making it difficult to digest the information or understand what the intended key takeaways are from the plots. Please consider reducing the number of subplots and more explicitly stating what trends the reader should notice.
This is a good point. We have now included subfigure reference labels and more explicitly described what each show.

6. Figures 10 and 11 show that the wind speed and power ratio of the model does not pass through zero for some cases as would be expected. Additionally, since the model is not tested on an unseen holdout test dataset it is not possible to tell if the NTFs represent meaningful and repeatable empirical relationships or overfitting to the data.
We agree that these observations demonstrate some of the limitations of the approach and/or available data. The discussion provides some examples of possible causes. As discussed above, the k-fold model validation approach demonstrates that the model did not overfit the data.

Minor Comments
• Section 3.2 would benefit from explicit equations defining the filtering metrics. In general, the methods section is text-heavy and could benefit from more equations which concisely describe the methods introduced.

We have rewritten the section to improve readability.
[[Still need to add - Aidan will work on this.]]

• Figures 2 and 4 would be clearer if point density were shown directly (e.g. using a KDE plot).
Both of these plots have been updated as KDE plots as recommended.

• The terms “predicted” and “estimated” are ambiguous since there are two models in use: the neighbor-based yaw-aligned quantity estimator and the NTF model. It may be clearer to define variables such as the ratio of the yaw-misaligned wind speed to the yaw-aligned wind speed called a wind speed ratio. Then, the goal of the NTF is to predict the wind speed ratio. In Figure 10 you could then plot the wind speed ratio predicted by the NTF removing any ambiguity. Similar variables could be defined for the wind vane bias or power ratio. These quantities could be defined in the methods section and used throughout for consistency and clarity.

This is a complexity in this paper that is difficult to explain clearly since we are using the output of one analysis (consensus) as the input to the second analysis (NTF function fitting). We have included a terminology section to clarify our terminology and worked to ensure that it is used consistently throughout the paper.

• The titles of Sections 4.1 and 4.2 are not descriptive enough. Please use titles which better explain the content of the section such as “Estimating expected yaw-aligned operation using weighted neighbor averaging methods”.

This is a good recommendation, we have updated the section titles to clearly describe the contents in these sections.

• The use of the term “stable” to mean a lack of spatial variability is confusing (L212, L333) since stability is commonly used to describe thermal stratification of the atmospheric boundary layer

We have corrected this terminology confusion and refer to the goal to filter data with high spatial variability as filtering for uniform spatial conditions.

• The acronym CRMSE should be defined when first introduced (Table 5).

This has been corrected.

• Table 5 would be more informative if measures of variability and uncertainty (e.g., standard deviations, standard errors) were included. Is the difference between methods statistically significant?

While the difference between the methods is relatively small the point is to select the method with the smallest error in an optimization sense and further statistical analysis of these errors was not conducted.

• For conciseness and clarity, I prefer to have all subfigures individually lettered, and then to use the lettering in the captions (e.g. Figure 5) to refer to the subfigures. Also, in general all references to the figures in the text should have the figure number and the letter of the subfigure (e.g. L271).

We have updated our subfigures to include letters and references to these in the captions and text as appropriate.