Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Measuring wake deflection from SCADA data during wake steering using machine learning

To cite this article: Nathan Post *et al* 2024 *J. Phys.: Conf. Ser.* **2767** 042031

Journal of Physics: Conference Series **2767** (2024) 042031 doi:10.1088/1742-6596/2767/4/042031

Measuring wake deflection from SCADA data during wake steering using machine learning

Nathan Post, Cheng Zheng, Daniel Kerrigan†, Enrico Bertini†, Melanie Tory

Roux Institute, Northeastern University, Portland, Maine, USA
†Northeastern University, Boston, Massachusetts, USA

E-mail: `n.post@northeastern.edu`

Abstract. As wake steering is implemented on large wind plants, methods for evaluating how well these systems optimize plant performance are needed. The data analysis from experiments conducted so far generally focuses on evaluating the improvement in power or energy production of the plant and comparing that to the predicted improvement. In this study, we explore a new approach to provide additional insight and validation of optimization tools by measuring the wake deflection experimentally using only the downstream turbine as a sensor. This approach is demonstrated using the SMARTEOLE wind plant wake steering experimental data. Light gradient boosting machines (LGBM) and Gaussian process (GP) machine learning models are trained and then interrogated by making a set of predictions under defined conditions to determine the wake deflection observed at the downstream turbine. The data set is sparse, particularly for larger yaw angles, and noisy due to factors not captured in the SCADA leading to relatively high uncertainty in the predictions. The GP model is better suited to smoothly fit this data. Using the GP model, a wake deflection of 0.35 rotor diameters at 3.7 rotor diameters downwind is estimated for a yaw error of 20 degrees on the upwind turbine.

1. Introduction

Wind farm flow control (WFFC) is gaining increasing interest in the industry with the goal of reducing wake interaction between turbines and improving overall plant performance. One implementation of WFFC is wake steering in which upwind turbines are intentionally yawed so that their rotors are not perpendicular to the wind in order to redirect the wakes away from downwind turbines. The control decision to off yaw a turbine is typically based on optimizing the wind plant performance in a tuned phenomenological model such as Flow Redirection and Induction in Steady State (FLORIS) [1]. In some implementations a table of yaw offsets is generated as a function of wind direction and wind speed from this optimization. This table is then used in the control of individual turbines. Because WFFC implementations rely on using a model to optimize the plant the result is an open loop controller and the desired performance improvement will only be possible if the model is accurately representing the real flow through the plant. Thus, model validation across a wide range of environmental conditions and plant layouts is required.

When comparing phenomenological model results with experimental data on actual wind plants, prior studies have focused on evaluating the improvement in farm power in specific conditions or total energy production and comparing that to the FLORIS model predictions for

Journal of Physics: Conference Series **2767** (2024) 042031 doi:10.1088/1742-6596/2767/4/042031

the site. Often this is done by comparing the power ratio between a waked turbine and unwaked turbine(s) referred to as the reference turbine(s). This power ratio is then binned and averaged as a function of wind direction with the wake controller on and off [2, 3]. While such an analysis provides information on how well the energy improvement is being captured by the model, it does not provide a specific indication of how the wakes are changed compared to predictions in different environmental conditions. The binning of data obscures the combined influence of other variables including wind speed and turbulence intensity on these measurements. Breaking the bins into further dimensions reduces the data available in each bin, increasing the uncertainty. To measure wake steering effects, experiments leveraging scanning LiDAR measurements have been conducted [4, 5]. These studies have been used to validate Large Eddy Simulations (LES) models of wake position downstream of a turbine that is at a yaw angle to the incoming free flow under specific sets of inflow conditions. LES models have in turn been used to validate predictions of phenomenological models (e.g. FLORIS) in specific conditions on small numbers of turbines with moderate 3 to 6 rotor diameter spacing. However, these techniques are not scalable to validating model prediction of wake effects across many turbines or deep in a large plant as will be the case for commercial installations.

Here we describe the initial development of a technique to evaluate wake modification by wake steering using trained machine learning (ML) models to predict power ratios. These models are trained using turbine supervisory control and data acquisition (SCADA) data collected from the turbine controllers during a wake steering experiment. We then interrogate the machine learning model to evaluate the modification of the wake due to wake steering using the downwind turbine predicted response as a sensor. This approach has the advantage that it has the potential to be applied to large wind plants to identify wake deflections without the use of LiDAR or other expensive independent measurements. Because environmental conditions are continuously changing when running a wake steering experiment, it is not possible to only change one input variable (feature) at a time while holding other features constant to determine influence of that one feature. However, if ML model has learned the behavior of the physical system from the available data then it is then possible to make predictions for specific feature combinations. By selecting specific values for most features and varying one feature at a time, we can understand the effects of individual features on the wake steering result. The objective in this study is to use the downwind turbine as a sensor and to apply machine learning models to measure how far wakes are deflected at specific environmental conditions (e.g. wind speed and turbulence intensity) as a function of the yaw misalignment of the upwind turbine. A more in-depth and broader understanding of wake deflection generated in this way could better inform the optimization of wake steering controllers.

2. Methodology

The power ratio of the downwind turbine relative to an undisturbed reference turbine or turbines provides a measure of the *wake intensity* experienced at that location and time. By training a machine learning (ML) model on the available data (features) to predict the power ratio (target), it is then possible to explore how the wake intensity varies as a function of wind direction and yaw misalignment at a given wind speed and turbulence intensity (Ti). This virtual experiment is shown conceptually with two different wind directions in Figure 1. Changing the wind direction is equivalent rotating the wind plant coordinate system and thus moving the downwind turbine across the wake of the upstream turbine. The center of the wake is determined by the wind direction that gives the minimum power ratio. As the yaw angle of the upstream turbine is adjusted, we expect the wake to deflect and wind direction where the minimum power ratio occurs to change. The corresponding wake deflection in rotor diameters is calculated from the change in wind direction as

$$\text{wake deflection} = \text{distance between turbines} \times \sin(\alpha) \tag{1}$$

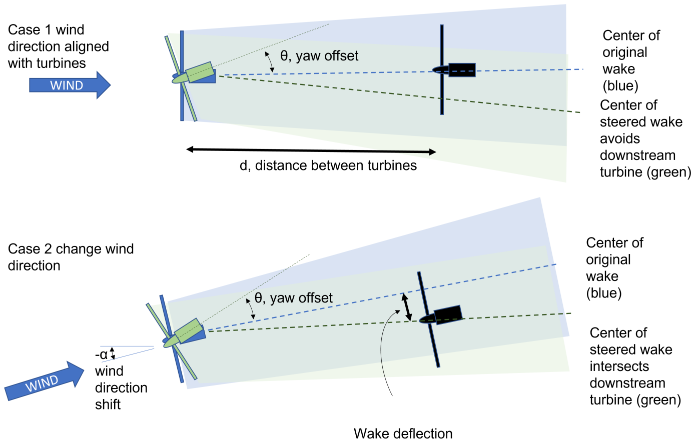


Figure 1: Conceptual experiment where the wind direction is changed by angle α for a given upstream turbine yaw misalignment (vane angle) until the downstream turbine is again in the maximum wake

where α is the difference in wind direction corresponding to the minimum power ratio relative to the wind direction minimizing the power ratio without a yaw offset. Although it is not possible to do a controlled experiment like this because we don't have control over the wind direction on a real wind plant, if the ML model is able to learn the response of the turbine system to these variables from the data, we can then run the experiment using predictions from that model trained on the real data.

3. Data preparation and model training

3.1. Data preparation

The proposed approach is demonstrated using a simple wake steering data set provided by the SMARTEOLE project [6] and described in [7, 8, 9]. This data set was collected on a 7-turbine plant and it includes 1 minute SCADA data for about 3 months of wake steering where turbine SMV6 was actively steered. The wind plant layout is shown in Figure 2.

The SCADA data were filtered and pre-processed and northing calibration was accomplished following a similar procedure as described by Simley et al. [7]. The data processing steps used code adapted from the FLASC [10] example code for this wind plant. Data from reference turbines SMV7, SMV3, SMV2 and SMV1 are averaged to calculate the reference power, reference wind speed, reference wind direction, and reference turbulence intensity (Ti). When calculating these averages, only turbines that are operating normally (producing power, not derated) are included in the calculation of reference values. The Ti for each reference turbine is calculated as the standard deviation of wind speed divided by the average wind speed as recorded by the nacelle anemometer of that turbine on a 1-minute basis. Then the Ti values for each valid reference turbine are averaged together to calculated the reference Ti. The data were further filtered to periods when SMV6 and SMV5 were operating normally and when the reference wind direction was between 195 and 241 degrees to focus only on data where significant wake interaction between SMV6 and SMV5 is present. After filtering 20974 samples of 1 min data remained for training the model. For the selected wind direction range, the reference turbines should all be unwaked and thus the reference values are a spatial average across the plant and provide an estimate of what the conditions and performance at the test turbines SMV6 and SMV5 would be if flow control was not active. The SMARTEOLE data set included a toggling

Journal of Physics: Conference Series **2767** (2024) 042031 doi:10.1088/1742-6596/2767/4/042031

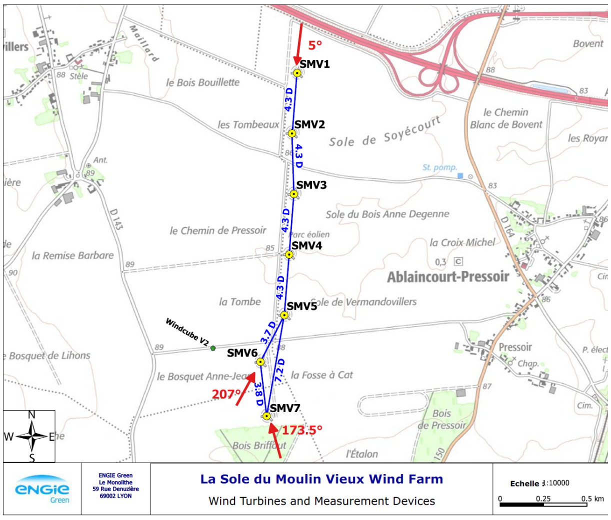


Figure 2: Layout of the wind plant and experimental setup for wake steering experiment [7].

period of 60 minutes where flow control was toggled on and off creating sequential alternating flow control and baseline periods. Because we are training a ML model to predict the waked turbine performance, all data from both flow control and baseline periods could be used in training and validating the model.

3.2. Selecting the target

It is possible to train a model to predict the downstream turbine power in kW directly and then calculate the power ratio in post-processing. However, this leads to a model that predicts the power curve of the turbine where the reference wind speed is the most important feature. Such a model does not do a good job of detecting the wake interaction effects because small changes in reference wind speed will have a large effect on power compared to the magnitude of changes in the wake through wake steering. By selecting the target as the ratio of SMV5 power to the reference power, wind direction becomes the dominate feature and wake effects are emphasized compared to changes in wind speed. The power ratio is calculated as:

$$\text{power ratio} = \frac{\text{power SMV5}}{\text{reference power}} \tag{2}$$

However, using the power ratio as the target results in an asymmetric error function (difference between the prediction and the target) for ML model training because values of the target where the power of SMV5 is less than the reference power will be between 0 and 1 while those where the power of SMV5 are greater than the reference power will be between 1 and ∞. Instead the log of the power ratio, log(power ratio) is taken as the target so that ratios above and below 1 are weighted equally depending on how far the SMV5 power is from the reference. The log power ratio is used as the target in training all models and shown in the predicted results.

3.3. Determining the features

Features representing the environmental conditions and the yaw position of the upstream turbine relative the wind in the SCADA data are considered. The use of features that are already included in the target, for example the reference power or power produced by individual turbines should be avoided as the model will focus on this highly correlated feature and not pick up on the specific changes in wake interaction. Data coming from other sources, for example LiDAR

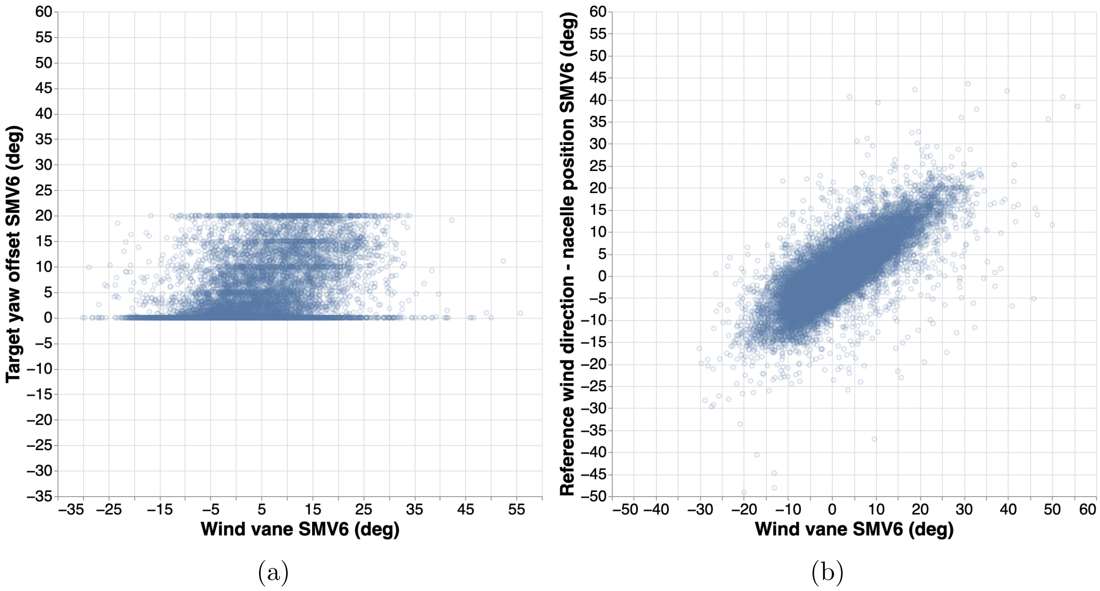


Figure 3: Comparison of potential features to represent the yaw offset of turbine SMV6. a) Target yaw offset vs. wind vane angle for SMV6. b) Wind direction - nacelle position of SMV6 vs. wind vane angle for SMV6.

Table 1: Nested cross-validation results for the LightGBM modeling process with different yaw position features. The results when only using the wind direction, wind speed, and turbulence intensity features and not yaw position are provided for comparison.

Yaw position feature	Mean of RMSE	Standard Deviation of RMSE
Wind vane of SMV6	0.342	0.187
Target yaw offset of SMV6	0.342	0.192
Wind direction − nacelle position of SMV6	0.348	0.190
None	0.347	0.188

or MET tower data are not included in this study as we want to focus on what we can do with only SCADA data. While individual measurements from undisturbed wind turbines can be used as features, the reference measurements provide some spatial filtering and allow some of the individual turbines to be offline and excluded for some of the data points. Thus, the environmental condition features used are the reference wind speed, reference wind direction, and reference Ti. In addition, a measure of the yaw offset of the upstream turbine is required. Three methods are considered for evaluating this yaw position of turbine SMV6: 1) using the vane angle measurement directly, 2) using the target yaw offset assigned by the wake steering controller, and 3) taking the circular difference of the reference wind direction and the turbine nacelle position. The relationship between the vane angle and these two alternative measures is shown in Figure 3. To identify the most useful yaw position feature, we trained light gradient boosting machine models using each feature in turn and we found that the selection of the position feature has a very small impact on the the cross-validation root mean square error (RMSE). Results of the feature selection study are summarized in Table 1. Given that wind vane angle has higher variance and is a directly measured feature, trends of wake steering using wind vane angles for yaw misalignment are potentially the most meaningful and thus it is used for the subsequent predictions.

3.4. Machine learning models

Two types of machine learning models are explored: a light gradient boosting machine (LightGBM) and a gaussian process (GP). For each type of model, the hyperparameters are tuned to minimize the RMSE in a cross validation.

The LightGBM library [11] was used to train light gradient boosting machines. These are decision tree-based models. We performed nested cross-validation to evaluate our modeling process and choice of yaw position feature, with 10 folds in the outer cross-validation and 5 folds in the inner cross-validation. Wind plant data is highly auto-correlated and thus if a random train/test split is used, the test and training data points will be nearly identical giving an overestimation of the model performance. To avoid this the data set was separated in time for each test/train split so that the model is exposed to new data to predict for the testing portion. In this case, the testing data for a given fold consisted of data points that were in sequence directly following the training data. In the inner-cross validation, we used Optuna [12] to sequentially tune the models' hyperparameters [13]. The mean and standard deviation of the RMSE scores of the outer 10 folds are reported in Table 1 for each set of features. To train our final model the best hyperparameters from the cross validation were selected and then used those parameters are used to train a model on the entire training dataset for subsequent predictions.

Gaussian process regression is a non-parametric Bayesian approach. A GP uses a prior function and its derived Gaussian distributions to infer the posterior function which will be used to make predictions at new data points. The fundamental principle of a Gaussian process states that the function values at point x follows a joint probability distribution that is a multivariant Gaussian, given as:

$$p(f|X) \sim \mathcal{N}(f|\mu, K) \tag{3}$$

where $f = (f(x_1), ..., f(x_N))$, $\mu = (m(x_1), ..., m(x_N))$ and $K_{ij} = k(x_i, x_j)$, m is a *mean* function, and k is a positive definite function, also known as *kernel* or *covariance* functions. The mean will be determined in the learning process, hence a Gaussian process is mainly defined by the kernel K. By assuming that the objective function has an independently and identically distributed noise given by σ, then the posterior of new function values f' given previous observations (X, y) and new data points X' can be written as:

$$p(f'|X', X, y) \sim \mathcal{N}(y'|\mu', \Sigma' + \sigma^2 I) \tag{4}$$

The mean prediction μ' is a linear combination of observations y, and the new covariance matrix $\Sigma' + \sigma^2 I$ is associated with covariance matrix of observed data and new data. Gaussian process is versatile and robust for making predictions in that it interpolates among observations, the predictions come together with uncertainty estimates and it can use different kernels to fit the data.

However, since a Gaussian process model is related to the inverse of high-dimensional matrix, it comes with a cubic scaling ($\mathcal{O}(n^3)$) with the number of data points (n). To train a good Gaussian process regressor with limited computational resources, we first downsampled the dataset with the objective of balancing the amount of data based on the feature 'target yaw offset of turbine SMV6'. The reduced training dataset is created by taking all of the data with a target yaw offset of 5 degrees or more and then including an equal number of data points randomly sampled from the data with a target yaw offset of less than 5 degrees. Then we trained a Gaussian process model using the 80% of the reduced data and we tested the model on the remaining 20% data. The data was split sequentially based on time without shuffling. The Gaussian model is then trained using a radial basis function as the kernel for which the length-scale parameters are optimized during fitting. The noise kernel with a variance of $\sigma = 0.5$ was determined by a hyperparameter tuning process where we select a handful of σ values and choose the value that gives the smallest RMSE on a hold-out set.

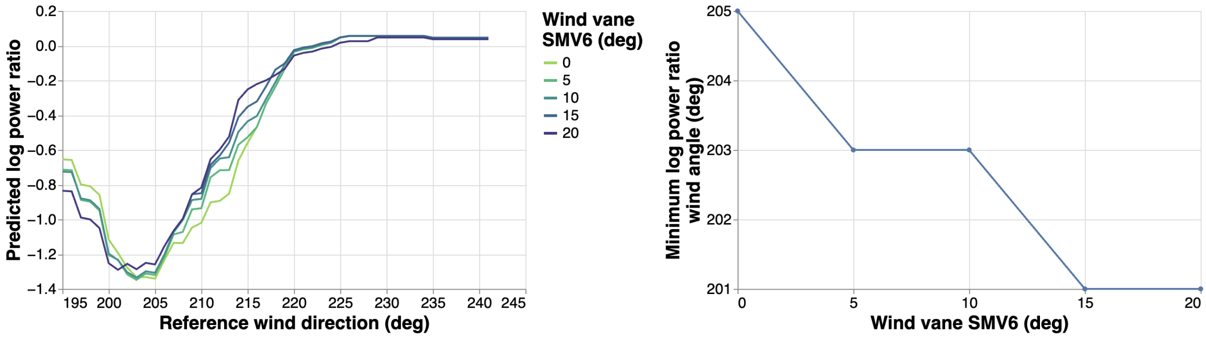


Figure 4: Left: Predicted log power ratios for a light gradient boosting machine. Wind speed is held at a constant 8m/s and turbulence intensity is held at a constant 0.10. Wind direction is varied from 195-241 degrees. The wind vane of SMV6 is varied from 0 to 20 degrees. Right: The wind direction at which each value for the wind vane of SMV6 resulted in the lowest predicted log power ratio.

4. Results and discussion

The trained models are used to explore the wake behavior as a function of SMV6 wind vane (yaw misalignment). This can be done for any set of environmental conditions (wind speed and turbulence intensity) that are well represented in the training data. In this case, we present results with a wind speed = 8 m/s and Ti = 0.1. The specific values of wind speed and Ti were selected because they are in the mid-part of region 2 where waking has a significant impact on power production and because they represent feature regions where we have many data points hence we expect the machine learning models to do a decent job in the given feature space.

4.1. Light gradient boosting machine results

Figure 4 (Left) shows the model's predicted log power ratios for various wind directions and wind vanes when holding the wind speed and turbulence intensity constant, as described above. Figure 4 (Right) plots the wind direction that minimizes the predicted log power ratio for wind vane values between 0 and 20 degrees. The LGBM predictions do not provide a smooth curve through the wake region and thus there is uncertainty in the wind direction corresponding to the minimum target value predicted. Despite this, we do see a trend with the wake predicted to shift by about 4 degrees under a yaw angle of 15 to 20 degrees.

Figure 5 shows the feature importance scores for the model. It is interesting to note the relative low importance of the wind vane feature, which is also reflected in Table 1, which shows that there is only a very small improvement in RMSE when using the wind vane of SMV6 over not using any measure of yaw position. Another notable aspect of the results in Table 1 are the high standard deviations of the RMSE scores across the outer 10 folds. This is an indication that the change in wake behavior as a function of vane angle is small and it is hard for the model to distinguish it from the noise in the data. Figure 6 uses box plots to visualize the distribution of wind speeds in the test sets for each fold of the outer cross-validation. Each box plot is colored by the RMSE of its fold. The RMSE is notably higher for folds 5, 9, and 10, which also correspond to the folds with the lowest wind speeds on average. A longer duration of data than was available would help to reduce model uncertainty by ensuring more periods with sufficient wind speeds.

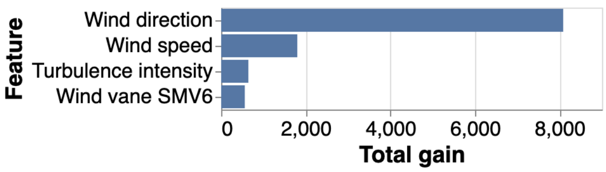


Figure 5: The LightGBM model's feature importance scores.

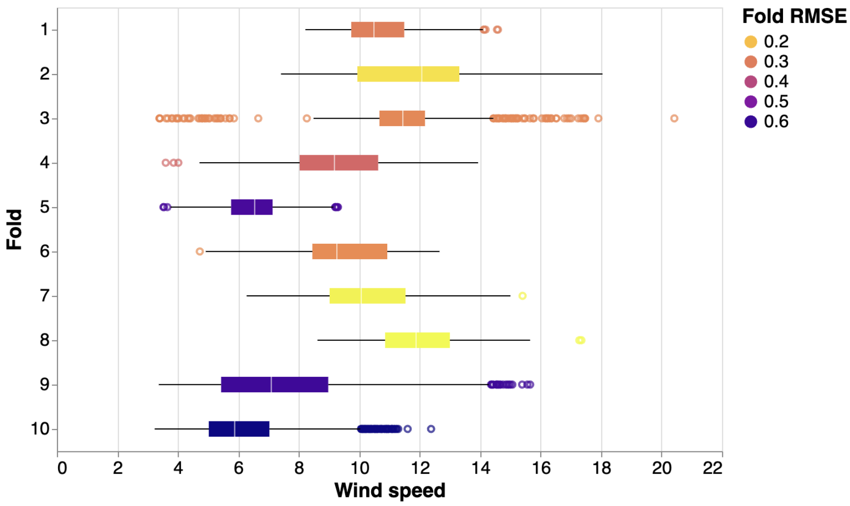


Figure 6: The relationship between wind speed and RMSE for the test sets of the outer cross-validation folds when using the wind vane of SMV6 feature. The folds whose test sets have the lowest average wind speeds tend to have the highest RMSE.

4.2. Gaussian process model results

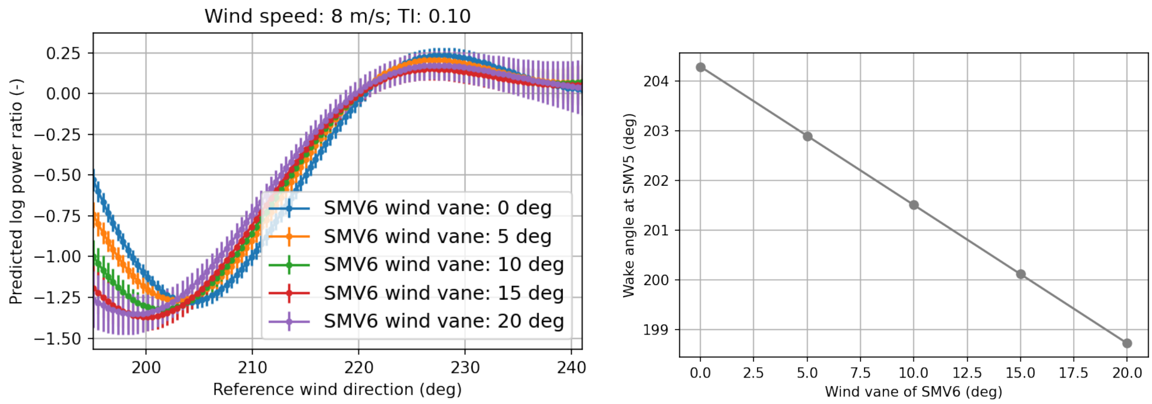
Using the trained GP model, the predicted mean of log power ratios and corresponding error bars *versus* wind directions are shown in Figure 7a, for the selected wind speed of 8.0 m/s and turbulence intensity of 0.1. Each curve represents results at a different wind vane angle of SMV6. The uncertainties for non-zero wind vanes are much larger than that for zero wind vane predictions because the data is sparse for high vane angles (see Figure 3a) and intentionally yaw offsets were not applied for all wind directions in the data set. The Gaussian process model predicts a nearly linear relationship between the vane angle of SMV6 and the offset angle of the wake as shown in Figure 7b indicating negligible higher order effects in this application. The fitted linear function for the wake angle measured at SMV5 (α_{wake}) *verses* the wind vane angles of SMV6 (θ_{wv}) as:

$$\alpha_{\text{wake}} = 204.3 - 0.278 * \theta_{\text{wv}} \tag{5}$$

Thus, one degree of wake angle shift requires around 3.6 degrees of wind vane angle increase.

4.3. Wake deflection estimates

Finally, using Equation 1 we calculate the corresponding deflection of the wake at turbine SMV5 3.7 rotor diameters down wind as a function of the upwind SMV6 turbine vane angle. These results are shown in 8. A maximum deflection of 0.35 rotor diameters is identified using the GP model while a smaller 0.25 rotor diameters is predicted with the LGBM model.



(a)　　(b)

Figure 7: Gaussian process ML predicted log power ratio SMV5/SMV7 at fixed wind speed: 8.0 m/s and Ti: 0.10. (a) and wake angle of minimum log power ratio (b)

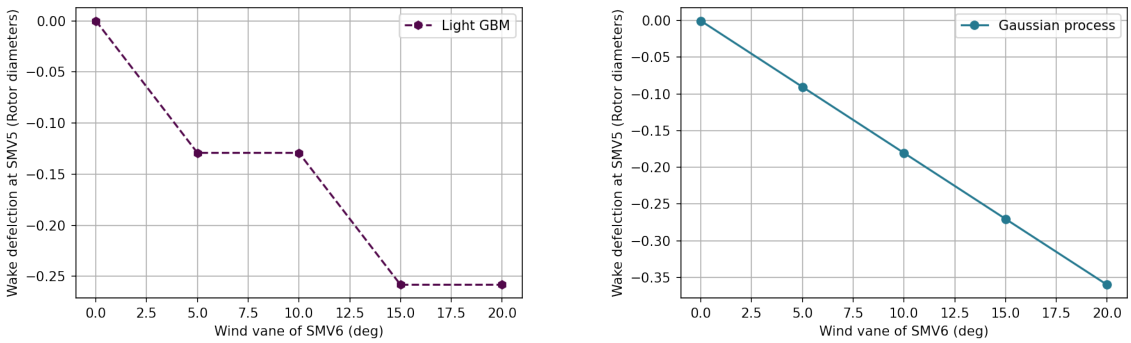


Figure 8: Wake deflection at SMV5 as a function Wind vane angles at SMV6 for two types of models: a light gradient boosting model (Left) and a Gaussian process model (Right).

5. Conclusion

In this paper a new approach was demonstrated to use machine learning models trained on SCADA wake steering experimental data to measure the effect of wake steering on the wake position by treating the downwind turbine as a sensor. While the approach shows some promise, the results should be evaluated carefully in light of the uncertainty due to several factors, including that the feature importance of the upwind turbine vane angle is small in the LGBM model and that the differences in target impacted by upwind turbine vane angle, which we aim to detect, are small relative to the overall noise in both the features and target. Furthermore, the intentional yaw offsets of the turbine are highly correlated to the wind speed and direction because SCADA wake steering experiments attempted to optimize plant performance and thus the data does not incorporate a wide range of yaw offsets for all conditions under consideration. This leads to a sparse data set which increases the uncertainty of model predictions, particularly for large yaw offsets where there is relatively little data. The sparse training data also suggests that one should carefully examine the model predictions in extrapolated regions which are not well represented in the data set. The data that is available in many cases is an “accident” as a result of wind direction shifts that the turbines have not yet adjusted for and thus in this static modeling approach where each minute of data is taken as a unique data point, dynamic situations lead to increased data noise. Future work will incorporate recent history into the feature set in order to enable the models to capture the plant dynamics better.

Despite these shortcomings in this initial study, an estimate of the wake deflection was made and for 8 m/s with a 10% turbulence intensity a wake deflection of 0.25 to 0.35 rotor diameters was estimated at 3.7 rotor diameters downstream for a yaw offset of 20 degrees. Future work will include expanding these prediction results to other environmental conditions and comparing them to wake deflection model predictions. Analysis of data from other wind plants and wake steering experiments will be useful in further assessing this approach to evaluating wake steering. If successful, these types of experimental estimates can be compared to FLORIS simulation results as part of an effort to further validate and refine or tune these models. Better model training will be possible using experimental data where the yaw offsets applied are more randomly distributed rather than being a function of wind speed and direction although that is obviously less ideal for long-term operation of a plant. Expanding the approach to the interaction of multiple turbines and multiple rows of turbines is also of interest as is exploring data sets with different wind turbine spacing.

Acknowledgements
The dataset used in this research has been obtained by ENGIE Green in the scope of French national project SMARTEOLE (grant no. ANR-14-CE05-0034).

References

[1] 2024 NREL/floris National Renewable Energy Laboratory URL `https://github.com/NREL/floris`
[2] Fleming P, King J, Dykes K, Simley E, Roadman J, Scholbrock A, Murphy P, Lundquist J K, Moriarty P, Fleming K, van Dam J, Bay C, Mudafort R, Lopez H, Skopek J, Scott M, Ryan B, Guernsey C and Brake D 2019 *Wind Energy Science* **4** 273–285 URL `https://wes.copernicus.org/articles/4/273/2019/`
[3] Doekemeijer B M, Kern S, Maturu S, Kanev S, Salbert B, Schreiber J, Campagnolo F, Bottasso C L, Schuler S, Wilts F, Neumann T, Potenza G, Calabretta F, Fioretti F and van Wingerden J W 2021 *Wind Energy Science* **6** 159–176 URL `https://wes.copernicus.org/articles/6/159/2021/`
[4] Brugger P, Debnath M, Scholbrock A, Fleming P, Moriarty P, Simley E, Jager D, Roadman J, Murphy M, Zong H and Porté-Agel F 2020 *Wind Energy Science* **5** 1253–1272 URL `https://wes.copernicus.org/articles/5/1253/2020/`
[5] Fleming P, Annoni J, Scholbrock A, Quon E, Dana S, Schreck S, Raach S, Haizmann F and Schlipf D 2017 *Journal of Physics: Conference Series* **854** 012013 URL `https://dx.doi.org/10.1088/1742-6596/854/1/012013`
[6] Duc T and Simley E 2022 Smarteole wind farm control open dataset v1.0 Zendo.org URL `https://zenodo.org/records/7342466`
[7] Simley E, Fleming P, Girard N, Alloin L, Godefroy E and Duc T 2021 *Wind Energy Science* **6** 1427–1453 URL `https://wes.copernicus.org/articles/6/1427/2021/`
[8] Ahmad T, Coupiac O, Petit A, Guignard S, Girard N, Kazemtabrizi B and Matthews P C 2018 *IEEE Transactions on Sustainable Energy* **9** 1169–1176
[9] Macrì S, Duc T, Leroy A, Girard N and Aubrun S 2020 *Journal of Physics: Conference Series* **1618** 062058 URL `https://dx.doi.org/10.1088/1742-6596/1618/6/062058`
[10] 2023 NREL/flasc National Renewable Energy Laboratory URL `https://github.com/NREL/flasc`
[11] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y 2017 *Advances in Neural Information Processing Systems* vol 30 (Curran Associates, Inc.)
[12] Akiba T, Sano S, Yanase T, Ohta T and Koyama M 2019 *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* KDD '19 (New York, NY, USA: Association for Computing Machinery) pp 2623–2631 ISBN 978-1-4503-6201-6
[13] Ozaki K 2022 LightGBM Tuner: New Optuna Integration for Hyperparameter Optimization