Review of: Influence of environmental characteristics on power production in an offshore wind turbine in the Belgian North Sea

Summary

The article uses Shapley analysis to identify the key factors affecting power production in an offshore wind turbine operating under non-waked conditions. The study considers several factors, including wind speed, turbulence intensity, air density, wind shear, wind veer, and wind–wave misalignment. Some of these variables are obtained directly from SCADA and LiDAR datasets, while others are derived from ERA5 reanalysis data for the site.

The authors then train a gradient boosting model to predict the power deviation, defined as $(P_{\mathrm{SCADA}} - P_{\mathrm{powercurve}})$, using the calculated variables as input features. Subsequently, they compare the distribution of Shapley values across different operating regimes to identify the most influential factors affecting power production in each regime. Although the paper presents a novel and important methodological contribution to a relevant research topic, additional analysis and validation are required to substantiate the proposed methodology and support the conclusions drawn.

Therefore, I believe that the manuscript requires **major revisions** before it is suitable for publication.

Major Comments

Comment 1 – Lack of Feature Ablation Study

Issue: The manuscript relies exclusively on SHAP analysis to identify the factors influencing power production. However, SHAP values only quantify the contribution of each feature to the predictions of the trained model and do not establish whether a given feature is necessary for achieving good predictive performance. Consequently, features that are highly correlated with other inputs or act as proxy variables may receive high SHAP values despite contributing little additional predictive information.

Recommendation: The authors should perform a feature ablation (leave-one-feature-out) study by retraining the model after removing each input variable and quantifying the resulting degradation in predictive performance (e.g., RMSE, MAE, or R^2). Comparing the ablation results with the SHAP analysis would provide a more rigorous assessment of feature importance and help distinguish indispensable variables from redundant or highly correlated predictors. Such an analysis would substantially strengthen the validity of the conclusions regarding the factors affecting power production.

Comment 2 – Validation on a Single Turbine

Issue: The proposed methodology is developed and validated using data from a single wind turbine. While the results are promising, it is unclear whether the identified feature importance is specific to the analyzed turbine or is representative of offshore wind turbines more generally. Consequently, the current validation is insufficient to support broad conclusions regarding the factors influencing power production.

Recommendation: To demonstrate the robustness and generalizability of the proposed methodology, the analysis should be repeated for multiple turbines within the same wind farm. The authors should quantify the consistency of the resulting SHAP values across turbines, for example by reporting the mean and standard deviation of the feature importance rankings or other suitable statistical measures. If data from additional offshore wind farms are available, the methodology should also be validated across different sites to assess its transferability under varying environmental conditions, turbine layouts, and operating characteristics. Such an analysis would substantially strengthen the validity of the conclusions presented in the manuscript.

Comment 3 – Reliance on a Single Machine Learning Model

Issue: The SHAP analysis is performed exclusively using a gradient boosting model. Consequently, it is unclear whether the reported feature importance reflects the underlying relationships in the data or is influenced by the specific characteristics of the selected model. This limits the generalizability of the conclusions.

Recommendation: To demonstrate the robustness of the proposed methodology, the authors should repeat the analysis using additional machine learning models, such as XGBoost, Random Forests, and a neural network. Before comparing the resulting SHAP values, the predictive performance of each model should be evaluated using appropriate metrics (e.g., RMSE, MAE, or R^2) to ensure that the models achieve comparable levels of accuracy. Although the exact SHAP values are not expected to be identical across different models, the relative importance and ranking of the dominant features should remain qualitatively consistent. Such a comparison would provide stronger evidence that the reported conclusions are robust and not an artifact of the chosen learning algorithm.

Minor Comments

Comment 1 – Standardization for Linear Regression

To enable a meaningful comparison of the regression coefficients obtained from the multivariate linear regression model, the predictor variables should be standardized prior to model fitting. Without standardization, the magnitude of each coefficient is influenced by the scale and units of its corresponding variable, making direct comparisons misleading. Although standardizing the inputs will not improve the predictive accuracy of the model, it is necessary for interpreting the relative influence of the different predictors.

Comment 2 – Encoding of Atmospheric Stability

The manuscript states that the atmospheric stability classes were assigned numerical values because the qualitative classification is

> "For the SHAP analysis, each atmospheric stability class is given a numerical classification, as the qualitative classification is not interpretable in the SHAP process."

However, the adopted encoding strategy is not described or justified. The authors should clearly explain how the categorical stability classes were converted into numerical values (e.g., ordinal or one-hot encoding) and discuss the implications of this choice for both the machine learning model and the resulting SHAP analysis. This clarification is important, as different encoding strategies can influence both the learned relationships and the interpretation of feature importance.

Comment 3 – Sensitivity to the Definition of Turbulence Intensity

Issue: The manuscript discusses in detail the limitations of the conventional definition of turbulence intensity, $TI = \sigma_U/\overline{U}$, and argues that this metric may not adequately characterize the turbulent inflow conditions. However, despite this discussion, the subsequent analysis relies exclusively on this conventional definition. Furthermore, turbulence intensity is identified as one of the dominant features in the SHAP analysis, making it important to assess whether this conclusion depends on the chosen definition of TI.
Recommendation: The authors should investigate the sensitivity of their results to the calculation of turbulence intensity by considering alternative turbulence TI formulations available in the literature. The resulting feature importance should then be compared with that obtained using the conventional definition. Demonstrating that turbulence intensity remains an important predictor irrespective of the chosen formulation would substantially strengthen the robustness of the reported conclusions.