RC2: 'Comment on wes-2024-21', Anonymous Referee #2, 29 Apr 2024

"An interesting and novel approach is proposed for measuring turbine performance in time in the form of a virtual sensor signal called TPI"

- Thank you for your comments that are very much appreciated. We have adjusted the title and conclusion to include mention of a key contribution of the work being the introduction and use of TPI.

  Updated *title*:
  "Full-Scale Wind Turbine Performance Assessment using TPI: A Study of Aerodynamic Degradation and Operational Influences"

  Conclusion key contribution highlighted:
  "In response, a methodological framework was developed, integrating SCADA data with detailed O&M records to assess the impact of blade aerodynamic modifications. This approach aims to isolate the effects of these modifications amidst a multitude of factors influencing turbine performance. A key contribution of this work is the development of a controller-informed Turbine Performance Integral (TPI) method for the investigated turbine. Furthermore, STL is employed to further isolate long-term and trends and seasonal variations in performance.

  The proposed methodology focuses on the isolating the individual contributions of various factors to performance deviations. Notably, the efficacy of the TPI method is demonstrated by its ability to detect distinct seasonal variations in individual turbine performance without relying on direct wind speed measurements, comparison to other turbines, or the use of combined data sources."

1. There are numerous typos and grammatical errors that need to be fixed.

- Thank you for pointing this out. We have carefully reviewed and corrected these in the revised manuscript to ensure clarity and precision in our presentation throughout the document.

2. Citations are mixed with the body of text, leading to misunderstanding. I suggest wrapping them in brackets.

- Corrected throughout the document.

3. Line 188: The range between 20% and 37% seems quite specific. How was it defined? Visually or according to a max allowed change in pitch?

- Text improved for clarity:

  "Datasets of each turbine are loaded for the twelve year period. The integral or area under the generator speed curve between 20% and 37% of normalised power is monitored, a region empirically determined by observing linear segments of this curve - where the pitch angle is minimally active (see Figures 1 and 2). Buffers are added on each side of the range to accommodate transient behaviour."

4. Section 2.8 General comment: Brief explanations of the methods implemented, the context, and the reasons for the analyses are missing.

- Method section Improved:

  “Statistical analysis

  Moving beyond visual assessment and accounting for data uncertainty and limited sample sizes for certain categories, the application of statistical methods is imperative. While observational analysis provides valuable initial insights, it does not fully account for the intrinsic and nuanced variability within the dataset. The use of statistical tests provides a scientific methodology to determine the likelihood that observed effects genuinely exist and are not merely coincidental artefacts. This study aims to determine the significance of observed performance differences and ensure the reliability of the findings through appropriate statistical analysis. A sequence of statistical analysis is employed to gain an understanding of the dataset and the study's findings. First the normality of the 'difference' values across each event category is assessed to determine the appropriate statistical tests. Then, significance tests are conducted to evaluate whether the performance differences for each category are significantly different from zero. Finally, optimal sample sizes are considered through power analysis to ensure the reliability of findings. The following sections outline the rationale behind each statistical method.

  Normality tests are conducted to determine the distribution characteristics of the data, as this information is crucial for selecting appropriate statistical tests. For the normality assessment, the selection of suitable statistical tests for the analysis depends on the distribution characteristics of the 'Difference' values across each event category. The Shapiro-Wilk test (Shapiro255 and Wilk (1965)), known for its efficacy in assessing deviations from normality, was employed to test the null hypothesis that these values originate from a normally distributed population. It is important to note that this test can be sensitive to sample size, especially for very small or very large samples, with optimal performance for sample sizes around 20-50, its application requires careful consideration. A p-value threshold of 0.05 was used. Results above this threshold do not definitively prove normality but rather as an absence of sufficient evidence to the contrary, justifying a working assumption of normality.

  Significance test are carried out to determine whether the observed performance differences for each event category are statistically significant or merely due to chance. Significance tests were conducted to assess whether the performance differences for each event category were significantly different from zero. The choice of significance test was contingent upon the results of the normality assessment. Depending on the normality of the data in each category, as determined by the Shapiro-Wilk test, either parametric one-sample t-tests (Student (1908), for normally distributed data, or non-parametric Wilcoxon signed-rank tests were utilised. The latter being particularly valuable for non-normally distributed data (Wilcoxon (1945)), offering an alternative to parametric tests. This test has been previously employed in studies regarding condition monitoring and fault detection in wind turbines by Dao (2022).

  Sample size considerations are given to ensure that the study has sufficient statistical power to detect meaningful effects and to provide reliable results. The optimal sample size depends on a desired level of statistical power, the magnitude of the effect size to

be detected and the variability within the dataset. A power analysis (Cohen (1988)) was conducted to ascertain the minimum sample size necessary to detect a statistically significant effect with a specified confidence level. The desired power was set at 0.8, representing an 80% probability of correctly rejecting the null hypothesis when it indeed is false. The alpha, or significance level, was set to 0.05 for the alpha level, indicating a 5% risk of erroneously rejecting the null hypothesis when it is true. Particular attention was paid to categories with limited sample sizes (fewer than 20 data points), as these can reduce the power of statistical tests and increase the risk of errors. The effect size, indicative of the magnitude of the difference aimed to detect, was estimated from the current data set using Cohen's d as a metric, alongside measures of variability such as the standard deviation. It is important to acknowledge that using sample data as an estimate of the effect size and sample standard deviation as an estimate of the true population standard deviation is an approximation. The actual sample size required may vary considerably depending on the true effect size and population variability, which are unknown a priori. Therefore, the results of the power analysis should be interpreted as an initial guide, subject to refinement as more data becomes available."

5. Line 248 - "Normality Assessment": Ambiguous. More context is needed in terms of "what" and "why".

- Please see improved Statistical Analysis section above.

6. Line 389: How is the time window selected for the calculation of the gradient?

- The gradient was calculated between two points: the first two weeks prior to an event, accounting for repair time - and another three weeks after the event to capture potential trend deviations.