FEYZİYE SCHOOLS FOUNDATION

IŞIK UNIVERSITY

Statistical Analysis of the Relationship Between Wind Speed and Capacity Factor at Kelmarsh Wind Farm

Instructor: Dr. Habibe AKTAY

Bilgesu ÇAKIR, 22MISY1016

Can Deniz KOÇAK, 22MISY1026

Dec, 2025

Contents

Introduction and Literature Review

Identification of Specific Problem Area

The transition to renewable energy requires precise monitoring of power generation efficiency. A critical challenge in wind energy management is understanding the variability of power output in relation to meteorological conditions. Specifically, operators must understand how wind speed, the primary kinetic driver, translates into normalized efficiency (Capacity Factor) rather than just raw power output, to assess the reliability of specific wind farm sites.

Prevalence and Scope of Problem

The inability to accurately characterize the performance curve of wind turbines can lead to grid instability and inefficient energy dispatch. This problem is global in scope but requires site-specific analysis, as local wind profiles differ significantly. This study focuses on the operational scope of the Kelmarsh Wind Farm in Northamptonshire, UK. 52°24'05"N 0°56'35"W

Figure 1: Kelmarsh Wind Farm

Literature Review

In the past few years, researchers have increasingly leveraged machine learning to model the wind speed-power output relationship in wind farms. Bouyeddou et al. (2021) applied latent variable regression techniques, Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) to SCADA data from a 2.05 MW turbine. Their comparative study demonstrated that even relatively simple regression-based ML models can achieve high accuracy, with both PCR and PLSR yielding a coefficient of determination $R^2 \approx 0.93$ for wind power

prediction. This result indicates that a vast majority of the variance in power output (capacity factor proxy) can be explained by wind speed and a few principal components, reaffirming the dominant influence of wind speed on energy production even when using linear ML models.

More complex neural network architectures have also been explored to capture nonlinear and spatiotemporal patterns. Chen et al. (2021) developed a hybrid deep learning model combining a Long Short-Term Memory network (LSTM) with a Convolutional Neural Network (CNN) to predict turbine power outputs in a wind farm. The LSTM component modeled temporal dependencies in wind speed/power time series, while the CNN captured spatial correlations among multiple turbines. This joint LSTM+CNN model outperformed single-model baselines (including a standalone LSTM, a CNN, and a Support Vector Machine) in multi-turbine power forecasting, achieving the lowest prediction error (i.e. reduced RMSE and MAE) among the methods tested. By fusing time-series memory with spatial feature extraction, the study demonstrated improved predictive power, effectively mapping wind speed fluctuations to power output with higher fidelity than earlier approaches. This underscores how neural networks can learn the nonlinear wind speed-power curve (which is roughly cubic and saturating) more accurately and handle interactions between turbines (e.g. wake effects) in a farm.

Ensemble machine learning methods (which combine multiple decision trees or models) have shown strong performance as well. Bilgili and Gül (2024) evaluated several ML algorithms, including Decision Tree, Random Forest, k-Nearest Neighbors, and the gradient-boosted XGBoost, using real wind turbine operational data (SCADA measurements of wind conditions, turbine settings, etc.). Model performance was measured via common metrics (R^2, MAE, MSE, RMSE, MAPE), and XGBoost was found to outperform the other techniques, achieving the highest R^2 and lowest error overall. The superior accuracy of XGBoost was attributed to its ensemble learning mechanism, which better captured the complex relationship between wind speed and power output by iteratively reducing prediction errors. Likewise, Allal et al. (2025) reported that combining models in a hybrid framework yields significant gains in predictive accuracy. They introduced a hybrid approach that integrated time-series trend modeling (using Prophet) with ensemble regressors (CatBoost gradient boosting and Random Forest) for wind power forecasting across multiple horizons. In their experiments on a 2.5-year wind farm dataset, the hybrid model achieved root mean square error (RMSE) values of 30.6 (15-min ahead), 50 (1-day ahead), and 41 (1-week ahead), which represented at least a 50% reduction in error compared to the best single regression model. This highlights how ensemble and hybrid ML techniques can capture wind speed-power dynamics more robustly, improving short-term and medium-term predictions of capacity factor or output. The contribution of such studies is in demonstrating that combining multiple algorithms (or data-driven with time-series models) can account for both the non-linear

power curve and temporal variations, thus deepening our quantitative understanding of how wind speed fluctuations translate into power generation under various conditions.

Researchers have also started to focus explicitly on modeling capacity factor (a normalized efficiency metric) using ML. For example, Mathew et al. (2024) developed a one-dimensional CNN model with a "soft ordering" layer to estimate a wind farm's capacity factor over time. Their goal was to detect performance degradation with turbine age by tracking declines in capacity factor. Trained on multi-year data, the CNN-based model captured the underlying wind speed-energy output relationship with high precision, evidenced by a normalized RMSE of 0.102 and MAE of 0.035 on the test set. Such accuracy is notable given that capacity factor integrates both wind speed variability and turbine performance limits; the low errors suggest the model learned the site-specific power curve and seasonal wind patterns effectively. By applying neural networks to capacity factor directly, this study provided a novel way to assess operational efficiency trends, illustrating how ML can be used not just for short-term forecasting but also for diagnosing long-term wind farm performance relative to wind speed inputs.

Across these recent studies, machine learning models consistently confirm and quantify the strong relationship between wind speed and wind farm output. Techniques ranging from regression ensembles to deep neural networks have yielded high correlation coefficients or R^2 scores (often 0.8-0.95) when predicting power or capacity factor from wind speed and related features. Moreover, the inclusion of additional environmental and turbine parameters (e.g. wind direction, temperature, blade pitch), as well as advanced model features (like attention mechanisms or hybridizing physical models with ML), has further enhanced prediction fidelity in diverse settings (offshore, onshore, complex terrain, etc.). Each of these post-2020 contributions reinforces the understanding that wind speed is the key driver of capacity factor, while also demonstrating how modern ML tools can capture the nonlinear, site-specific nuances of that relationship. By improving predictive accuracy (lowering RMSE by substantial margins) and yielding high explanatory power, these studies collectively advance our ability to characterize and forecast wind farm efficiency using data-driven methods, a crucial step for optimized wind energy integration and planning.

Critique of Previous Research

A major limitation in the existing wind-energy analytics literature is its strong emphasis on predictive "black-box" modeling rather than interpretable, descriptive profiling of turbine performance. As your draft already notes, many studies prioritize complex AI forecasting pipelines for grid planning, but provide less insight into why or how efficiency varies at a site in a way that

is easily actionable for operators. This can weaken the operational interpretability of results, especially when the research goal is to characterize the wind speed-efficiency relationship rather than only forecast future outputs.

A second limitation is that many studies model raw power output (kW) rather than a normalized efficiency metric. Your paper correctly argues that using raw power makes it harder to compare performance across turbines with different rated capacities, whereas Capacity Factor offers a standardized dependent variable. As a result, the literature may overstate generalizability across turbine models and sites, because "high accuracy" in power prediction may partially reflect turbine rating and site-specific scaling rather than true efficiency dynamics.

Finally, many wind power ML studies implicitly assume that wind speed is the "dominant variable" and treat other drivers (air density, turbulence intensity, temperature, curtailment, icing, blade degradation, etc.) as secondary. Your methodology section acknowledges this exclusion explicitly as a limitation of the current scope. In practice, omitting these variables can leave a meaningful portion of capacity factor variance unexplained, and can also bias the estimated functional shape (e.g., saturation effects may be confounded with curtailment or downtime).

Gap in the Literature

Although recent studies increasingly apply machine learning to wind farm SCADA data, the majority of this work is oriented toward black-box power forecasting rather than interpretable, descriptive characterization of turbine performance. Moreover, many studies use raw power output (kW) as the dependent variable, which limits comparability across turbines and sites. As a result, there remains a clear gap in simplified, statistically interpretable analyses that focus explicitly on Capacity Factor as a standardized efficiency metric and quantify the strength and functional shape of the wind speed-efficiency relationship over long operational periods.

Purpose of Study and Research Questions

The purpose of this study is to statistically quantify the strength and shape of the relationship between wind speed and wind energy efficiency.

Research Question: To what extent does wind speed explain the variation in the Capacity Factor of turbines at the Kelmarsh Wind Farm?

Hypothesis 1 (H1): Wind speed has a strong, positive effect on the Capacity Factor.

Hypothesis 2 (H2): The relationship between wind speed and Capacity Factor is non-linear, following a sigmoid distribution where efficiency saturates at rated wind speeds.

Figure 2: Wind Generation & Capacity Factor Diagram

Methodology

Design

Type of Design: This study utilizes a Quantitative research design. The rationale is that the variables of interest (Wind Speed and Capacity Factor) are numerical (Ratio scale) and derived from objective measurement systems, requiring statistical analysis rather than qualitative interpretation.

Specific Design: The study employs a Descriptive and Correlational Design. The rationale is that the goal is to summarize the characteristics of the dataset (mean, variance, dispersion) and to test the strength of the association between two variables using bivariate analysis.

Specific Research Methods: The method used is Secondary Data Analysis. The rationale is that high-frequency historical operational data is already available via the Zenodo repository, eliminating the need for expensive and time-consuming primary data collection.

Sample and Procedures

1. Population and Sample:
    a. Population: The theoretical energy generation performance of all wind turbines at the Kelmarsh wind farm over their entire operational lifespan.

    b. Sample: The sample consists of historical SCADA data records spanning from 2016 to 2024 for the six turbines located at the Kelmarsh site.

2. Selection Procedures: The sample was selected using Non-probability Convenience Sampling. The data was chosen because it is a publicly available, open-access dataset provided by Cubico Sustainable Investments. Once selected, the data will be filtered to remove timestamps containing sensor errors or maintenance periods (where power output is zero despite wind presence).

Measurement

1. Instruments and Procedures:
    a. Instrument: The data was originally collected using the site's SCADA system.

    b. Independent Variable (Wind Speed): Measured in meters per second (m/s) using anemometers mounted on the turbine nacelles. This is a continuous, ratio-level variable.

    c. Dependent Variable (Capacity Factor): This variable measures efficiency. It is calculated using the formula:

$$\frac{\text{Actual Power Output}}{\text{Rated Peak Power}} x\ 100$$

    This is a continuous, ratio-level variable.

    d. Validity and Reliability: SCADA systems are the industry standard for wind farm monitoring. The sensors are calibrated instruments, ensuring high reliability. To ensure validity, we will follow the preprocessing steps outlined by Bouabdallaoui et al. (2025), removing non-physical values (e.g., negative power output).

Analysis Plan

The data analysis will proceed in the following steps using Microsoft Excel and Python:

1. Data Cleaning: Remove outliers and rows with missing values or negative readings.
2. Descriptive Statistics: Calculate the Mean, Median, Mode, Standard Deviation, and Coefficient of Variation (CV) for both Wind Speed and Capacity Factor to summarize the central tendency and dispersion.
3. Visualization: Generate Histograms to analyze the distribution shape (skewness) of the variables.
4. Bivariate Analysis: Create Scatter Plots to visualize the relationship between Wind Speed (X-axis) and Capacity Factor (Y-axis). Calculate the Pearson Correlation Coefficient (r) and the Coefficient of Determination (R^2) to test the hypotheses.

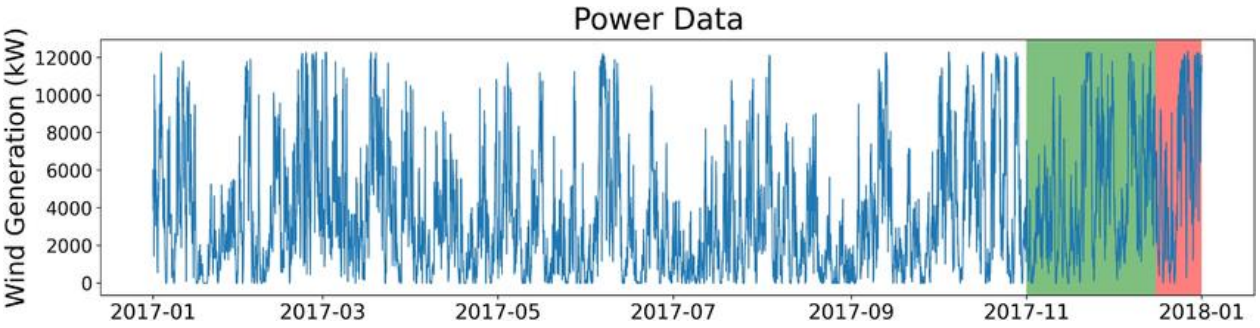

Figure 3: Power dataset for Kelmarsh wind farm (1-hour resolution). The training, validation, and test datasets are shown in white, green, and red, respectively.

Limitations

Sampling Limitation: Convenience sampling limits the generalizability of the findings. Results from Kelmarsh (an onshore site in the UK) may not apply to offshore sites or different climatic regions.

Variable Limitation: The study focuses only on wind speed. Other variables such as air density, temperature, and blade degradation over time are excluded from this specific analysis, which may leave some variance in the Capacity Factor unexplained. Although, more variables may be included in the research later.

References

Bouyeddou, B., Harrou, F., Saidi, A., & Sun, Y. (2021). *An effective wind power prediction using latent regression models*. In Proc. of the 2021 International Conference on ICT for Smart Society (ICISS) (pp. 1-6). IEEE.

Chen, X., Zhang, X., Dong, M., Huang, L., Guo, Y., & He, S. (2021). Deep learning-based prediction of wind power for multi-turbines in a wind farm. *Frontiers in Energy Research, 9*, 723775. https://doi.org/10.3389/fenrg.2021.723775

Bilgili, A., & Gül, K. (2024). Forecasting power generation of wind turbine with real-time data using machine learning algorithms. *Clean Technologies and Recycling, 4*(2), 108-124. https://doi.org/10.3934/ctr.2024006

Mathew, M. S., Kandukuri, S. T., & Omlin, C. W. (2024). Soft Ordering 1-D CNN to Estimate the Capacity Factor of Windfarms for Identifying the Age-Related Performance Degradation. *PHM Society European Conference, 8*(1), 9. https://doi.org/10.36001/phme.2024.v8i1.4028

Allal, Z., Noura, H. N., Salman, O., & Chahine, K. (2025). Machine learning-based prediction model of wind turbine power generation. *Cleaner Energy Systems, 12*, 100218. https://doi.org/10.1016/j.cles.2025.100218

Plumley, C., & Takeuchi, R. (2025). *Kelmarsh wind farm data* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16807551