# Explainable hybrid semi-parametric model for prediction of power generated by wind turbines

Alfonso Gijón[1,*][0000−0002−7400−8182], Simone Eiraudo[3,*][0000−0001−9831−2317], Antonio Manjavacas[1][0000−0001−7334−1074], Lorenzo Bottaccioli[3], Andrea Lanzini[3], Miguel Molina-Solana[1,2][0000−0001−5688−2039], and Juan Gómez-Romero[1][0000−0003−0439−3692]

[1] Dept. of Computer Science and AI, University of Granada, Spain
[2] Dept. of Computing, Imperial College London, United Kingdom
[3] Energy Center Lab, Politecnico di Torino, Italy
`* Corresponding authors: alfonso.gijon@ugr.es, simone.eiraudo@polito.it`

**Abstract.** The ever-growing sector of wind energy underscores the importance of optimizing turbine operations and ensuring their maintenance with early fault detection mechanisms. Existing empirical and physics-based models provide approximate predictions of the generated power as a function of the wind speed, but face limitations in capturing the non-linear and complex relationships between input variables and output power. Data-driven methods present new avenues for enhancing wind turbine modeling using large datasets, thereby improving accuracy and efficiency. In this study, we use a hybrid semi-parametric model to leverage the strengths of two distinct approaches in a dataset with four turbines of a wind farm. Our model comprises a physics-inspired submodel, which offers a reliable approximation of the power, combined with a non-parametric submodel to predict the residual component. This non-parametric submodel is fed with a broader set of variables, aiming to capture phenomena not addressed by the physics-based part. For explainability purposes, the influence of input features on the output of the residual submodel is analyzed using SHAP values. The proposed hybrid model finally yields a 35-40 % accuracy improvement in the prediction of power generation with respect to the physics-based model. At the same time, the explainability analysis, along with the physics grounding from the parametric submodel, ensure deep understanding of the analyzed problem. In the end, this investigation paves the way for assessing the impact, and thus the potential optimization, of several unmodeled independent variables on the power generated by wind turbines.

**Keywords:** Hybrid semi-parametric models · Explainable AI · Wind Turbines

## 1 Introduction

The growing use of renewable energies plays a pivotal role in tackling climate change and advancing towards a sustainable energy landscape. Concurrently,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_21

the rapid advancements in sensor and storage technologies have facilitated the accumulation of vast amounts of data, coupled with the rise of flexible and powerful data-driven and machine learning methodologies. In this context, the development of accurate and robust wind turbine (WT) models becomes essential for optimizing operations and automatic fault diagnosis.

The absence of precise and robust physics-based models for forecasting power production in utility-scale farms motivates the application of data-driven approaches. While neural networks are traditionally considered black-box models, the emergence of novel architectures capable of adhering to specific constraints, such as physics-informed neural networks (PINNs), improves their capabilities for modeling physical phenomena looking for accurate and robust models [5]. PINNs, although respecting certain physical constraints, still are non-explainable models with complex interpretation. Semi-parametric models are hybrid approaches bringing together physics-based and non-parametric methods. Indeed, they can provide high accuracy while preserving the interpretability of some modeled functional relationship [9].

This work focuses on the modeling of data from the four turbines located within the 'La Haute Borne' wind farm. Our main contribution lies in the development of a unified methodology aimed at effectively integrating physics-based and data-driven models within a common framework. Beyond the improvement in model accuracy, the explainability analysis of input feature importance through SHAP [8] values in the non-parametric submodel provides valuable insights into how input variables influence the output prediction.

This paper is organized as follows. An overview of wind turbine physics and modeling, alongside a description of the hybrid semi-parametric model are exposed in section 2. The main findings are reported and discussed in section 3, and finally, section 4 offers concluding remarks and outlines potential avenues for future research.

## 2 Computational methods

### 2.1 Physical background

Although challenging, the modeling of the low-scale aerodynamic behaviour of wind turbines can be achieved through physics-based fluid dynamics methodologies. Although the capacity of these models to predict the power generation of utility-scale wind farms is limited [7], partial physical information can be gleaned through the use of well-established equations relating certain high-scale variables. The power extracted by a WT from the kinetic energy of the incoming wind is given by:

$$P = \frac{1}{2} C_p \rho A v^3 \,, \tag{1}$$

where $C_p$ is the power coefficient, $\rho$ is the air density, $A$ is the area swept by the blades of the WT, and $v$ is the wind velocity.

The power generated by a WT is strongly related to the power coefficient, $C_p$, a dimensionless parameter accounting for nonlinearities and influenced by the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_21

inherent characteristics of the WT (such as its size, geometry and aerodynamic properties), as well as the operational conditions defined by variables such as wind speed and pitch angle. Typically, the objective is to optimize the value of $C_p$ to achieve maximal efficiency in converting wind energy into electrical energy, with a theoretical upper limit of 0.5926, known as the Betz limit [1]. In variable wind speed regions, the optimal power output of WTs is achieved through precise adjustment of the pitch angle, $\theta$, which is defined as the angle between the lateral axis of the blades and the direction of the relative wind. However, the complexity of pitch control for WTs stem from the inherent nonlinear dynamics of these systems and external disturbances. Several empirical formulas have been proposed to model the power coefficient [3,4], but they are not entirely satisfactory to model large amounts of data.

### 2.2 Hybrid semi-parametric model

The structure of hybrid semi-parametric models combines both parametric and non-parametric submodels, based on different sources of knowledge to construct comprehensive representations. In this work, non-parametric models are implemented as neural networks employing a multilayer perceptron architecture, with a flexible number of parameters that are not predetermined by prior knowledge. For our purposes, the prediction of the power is composed of two main components: a physics-inspired part, based on Equation 1, and a non-parametric part, aimed to predict the residues of the physics-inspired output with respect to the target variable, see Figure 1.

$$\hat{P} = P_{\mathrm{phys}}(\mathbf{x}) + P_{\mathrm{res}}(\tilde{\mathbf{x}}) \tag{2}$$

The physics-inspired submodel $P_{\mathrm{phys}}$ is driven by input variables that are readily interpretable and directly associated with the kinetic-electrical energy conversion process, i.e. wind velocity, pitch angle and rotor angular velocity. Besides, the residual submodel $P_{\mathrm{res}}$ uses a broader set of input variables extracted from the dataset, some easily interpretable and others difficult to interpret and incorporate into physical equations:

$$\mathbf{x} = (v, \theta, \omega)\,, \tag{3}$$

$$\tilde{\mathbf{x}} = (v, \theta, \omega, v_1, v_2, T_{\mathrm{out}}, T_{\mathrm{h}}, T_{\mathrm{r}}, T_{\mathrm{n}}, g_{\mathrm{v}}, g_{\mathrm{f}}, \alpha_{\mathrm{n}}, \alpha_{\mathrm{w}}, \alpha_{\mathrm{v}}, \alpha_{\mathrm{wc}}, \alpha_{\mathrm{nc}})\,. \tag{4}$$

From left to right, the variables defining $\tilde{\mathbf{x}}$ are: average wind velocity, pitch angle, rotor angular velocity, first anemometer velocity, second anemometer velocity, outdoor temperature, hub temperature, rotor temperature, nacelle temperature, grid voltage, grid frequency, nacelle angle, wind angle, vane angle, wind angle corrected, nacelle angle corrected.

It is important to note that the physics-inspired submodel, is trained with the power data as target, thereby providing a solid approximation of this quantity. Meanwhile, the residual submodel, further enhances the prediction of the power by integrating corrections derived from unknown physical factors, incorporating

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_21

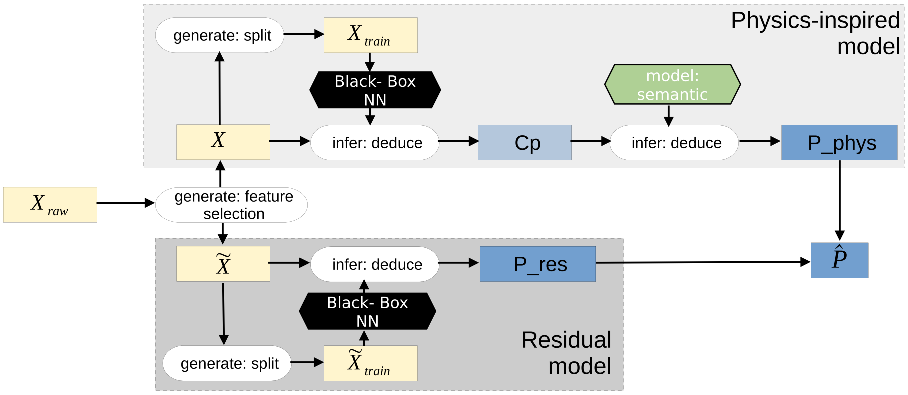


Fig. 1: Diagram of the hybrid model, according to the taxonomy presented in [2]. Rectangular, rounded, and hexagonal boxes represent data, functions, and models, respectively. Yellow and blue boxes are used for inputs and outputs, respectively.

variables that better describe the state of the wind turbine, such as the temperatures and orientation of its components. In the physics-inspired submodel, $C_p$ is predicted by a neural network and then the power is calculated using the semantic model provided by Equation 1. Then, the residue is calculated as $P - P_{\text{phys}}$ and used as input to train the residual submodel.

## 3 Results and discussion

Our experimental setup is based on artificial neural network models, trained using the *Tensorflow* library and hyperparameters optimized with the Hyperband algorithm from the *Keras Tuner* library. All the calculations were carried out in a computer equipped with an 11th Gen Intel Core i7-11800H processor, 16 GB RAM memory and NVIDIA GeForce RTX 3060 graphics card.

Before preprocessing, our dataset consisted of approximately 1 million instances. Initially, non-physical data points, such us those exhibiting a power coefficient higher than the theoretical Betz limit ($C_p > 0.5926$), were eliminated. The calculation of $C_p$ from direct measurements is sensitive to error propagation and only physically allowed values were preserved. Subsequently, we identified anomalous data by comparing the measured power with an estimation derived from the power curve using an iterative median technique. Any data point deviating from the median by more than 3 standard deviations ($3\sigma$) was deemed anomalous and removed. Additionally, a low-velocity power cutoff was applied to filter out noisy data at low velocities, where the relative error would have a more significant impact. Consequently, the dataset was reduced from 1 million to $7{\times}10^5$ data, accounting for 70% of the original data.

DOI: 10.1007/978-3-031-63775-9_21

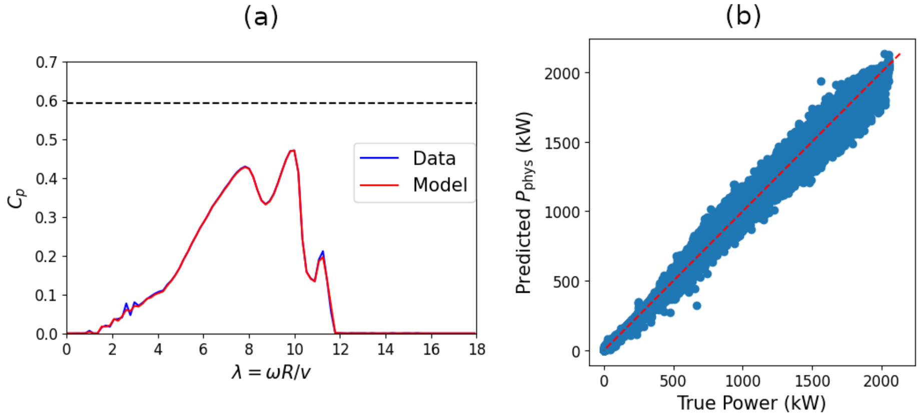


Fig. 2: (a) $C_p$ curve as a function of the tip speed ratio for the physics-inspired model. (b) Predicted vs true values for $P_{\text{phys}}$ computed through the intermediate power coefficient.

For training both submodels, the batch size was fixed to 128, employing the mean absolute error as the loss function, and reducing the learning rate when the loss stopped improving. The dataset was randomly split into 80 % for training and 20 % for testing. It was observed that 150 epochs were sufficient to reach a plateau in learning.

The hyperparameter search was carried out within the parameter space generated by the combinations of `n_layers` $\in \{1, 2, 4\}$, `n_neurons` $\in \{8, 16, 32, 64, 128\}$, `learning_rate` $\in \{0.01, 0.001, 0.0001\}$ and `activation_function` $\in \{$'relu', 'tanh'$\}$. For the physics-inspired submodel, $P_{\text{phys}}$, the hyperparameter search yielded a 2-layer architecture with 128 units per layer, a learning rate of 0.001, and ReLU as activation function. The regression of the intermediate variable $C_p$ with inputs $(v, \theta, \omega)$ is depicted in Figure 2(a) as a function of the tip speed ratio, $\lambda = \omega R/v$. A sigmoid output layer was employed to restrict the output within a fixed range of [0,1], which was subsequently converted to original units to respect the Betz constraint. Figure 2(b) illustrates the comparison between predicted and true values of $P_{\text{phys}}$ for the test dataset. The physics-inspired submodel achieves satisfactory performance by itself, demonstrating a mean absolute error (MAE) of 16.3 kW and a mean absolute percentage error (MAPE) of 3.71 %, as indicated in Table 1.

The neural network architecture utilized in the residual submodel, $P_{\text{res}}$, was identical to that of the physics-inspired model. In this case, the input comprises 16 variables used to make a prediction of the residual. As shown in Figure 3(a), the absolute residual power exhibits elevated levels within the medium power range (800-1800 kW), while remaining lower at both low and high power extremes. Despite being quite dispersed around the true value, the predicted residual power is adequately estimated, as represented in Figure 3(b).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_21

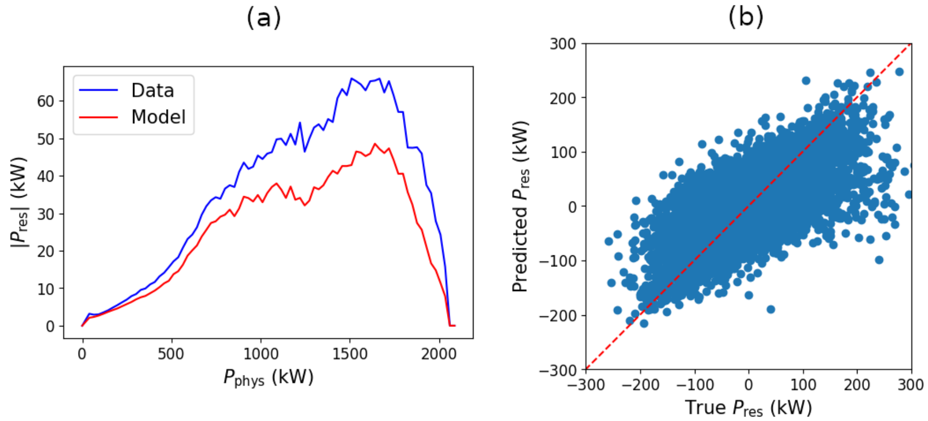


Fig. 3: (a) Absolute residual power as a function of the physical power for data and trained model. (b) Predicted vs true residual power for the test dataset.

| | PINN | Phys | Hybrid | Hybrid_CF |
|---|---|---|---|---|
| MAE (kW) | 15.55 | 16.31 | **10.51** | 10.89 |
| RMSE (kW) | 28.02 | 30.63 | **22.07** | 22.74 |
| MAPE (%) | 3.838 | 3.706 | **2.159** | 2.203 |
| R2 score | 0.9960 | 0.9953 | **0.9976** | 0.9974 |

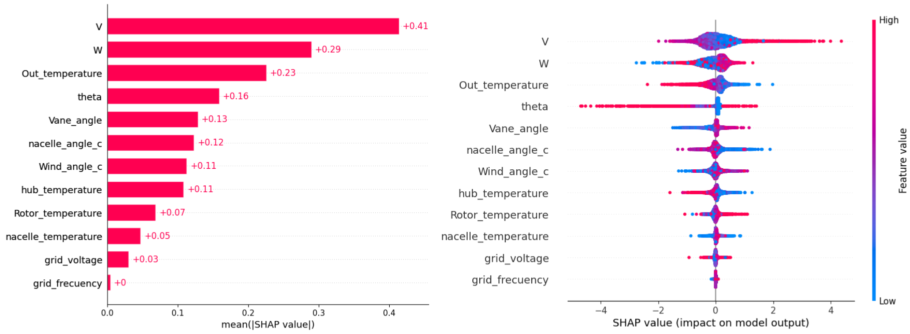
Table 1: Comparison of the performance metrics for different models. The PINN column is taken from reference [6].

The resulting hybrid model obtained by combining the physics-inspired and residual submodels surpasses the performance metrics to predict the generated power, achieving a MAE of approximately 10.5 kW and a MAPE of 2.16 %. As shown in Table 1, this represents an enhancement of approximately 35-40 % in the regression task, when compared to the physics-inspired submodel of this study or the physics-informed model of reference [6].

The physics-inspired part of our model is entirely interpretable, as it employs a physical equation and an intermediate variable constrained within a specific range. In contrast, the residual component operates as a black box, receiving 16 input variables and generating a prediction without explicit interpretability. As a first step in the explainability analysis, a linear correlation filter was applied to eliminate redundancies among the input variables. Specifically, those variables having a correlation coefficient exceeding 0.95 were excluded, resulting in a reduced set comprising 12 variables. The reduced hybrid model nearly achieves the same performance as the original one, as can be seen in Table 1.

To gain some insights into the relative importance of each input feature on the output, we show the mean absolute value of the SHAP values in Figure 4(a). As expected, the variables incorporated in the physics-inspired model ($v$, $\theta$, $w$) are the most influential. However, the direction of the wind, as well as the orientation

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_21



(a) (b)

Fig. 4: (a) Relative importance of each feature on the output. (b) Distribution of the impact of each feature on the model output. Color represents feature value.

of the nacelle, demonstrate considerable influence, motivating a deeper analysis and suggesting a potential optimization to increase electrical power generation under the same external wind speed and temperature conditions. An analysis of the impact distribution of each feature in Figure 4(b) reveals, for instance, that higher vane angle and wind angle contribute positively to power generation, while a higher nacelle angle and hub temperature exhibit a negative contribution.

## 4 Conclusions

In this study, we have designed and validated a hybrid semi-parametric model composed by a physics-inspired submodel and a non-parametric submodel, $P_{\text{phys}}$, for predicting the residual power of the physical term, $P_{\text{res}}$. The developed model, trained using real historical data of four turbines from a wind farm, results in an improvement of approximately 35-40 % in predicting the generated power. The physics-inspired submodel is inherently explainable due to its construction, leveraging a physical equation that relates the most critical variables of the system. However, the non-parametric residual submodel requires the analysis of SHAP values to comprehend the relative importance of the input features and their impact on the output power value. Our results suggest that certain angular variables could be adjusted to achieve higher power production.

It is noticeable that our methodology is versatile and can be applied to a wide range of problems where a physics-based model is available, offering approximate results, and additional data can be leveraged by a non-parametric data-driven submodel to predict the residual component and incorporate unknown physics.

Once deployed, this hybrid model could serve as an accurate regression-based anomaly detection method by comparing the deviation of new data from the model's prediction for a healthy state. All the models presented here are fully differentiable, enabling their utilization for developing optimal pitch angle controllers, thereby optimizing power generation across various wind speed regimes. While this hybrid model shows promise, further research is needed to asses their robustness across different turbines.

## Acknowledgements

This work was partially funded by the Spanish Ministry of Economic Affairs and Digital Transformation (NextGenerationEU funds) within the project IA4TES MIA.2021.M04.0008. It was also funded by ERDF/Junta de Andalucía (D3S project P21.00247, and SE2021 UGR IFMIF-DONES), and MICIU/AEI/ 10.13039/501100011033 and EU ERDF (SINERGY, PID2021.125537NA.I00). Acknowledgement is extended to ENGIE for providing such an interesting and well-documented sample dataset.

## References

1. Aerodynamics of Horizontal Axis Wind Turbines, chap. 3, pp. 39–136. John Wiley & Sons, Ltd (2011)
2. van Bekkum, M., de Boer, M., van Harmelen, F., Meyer-Vitali, A., Teije, A.t.: Modular design patterns for hybrid learning and reasoning systems. Applied Intelligence **51**(9), 6528–6546 (Sep 2021)
3. Carpintero-Renteria, M., Santos-Martin, D., Lent, A., Ramos, C.: Wind turbine power coefficient models based on neural networks and polynomial fitting. IET renewable power generation. **14**(11) (2020-08)
4. Castillo, O.C., Andrade, V.R., Rivas, J.J.R., González, R.O.: Comparison of power coefficients in wind turbines considering the tip speed ratio and blade pitch angle. Energies **16**(6) (2023)
5. Fernández de la Mata, F., Gijón, A., Molina-Solana, M., Gómez-Romero, J.: Physics-informed neural networks for data-driven simulation: Advantages, limitations, and opportunities. Physica A: Statistical Mechanics and its Applications **610**, 128415 (2023)
6. Gijón, A., Pujana-Goitia, A., Perea, E., Molina-Solana, M., Gómez-Romero, J.: Prediction of wind turbines power with physics-informed neural networks and evidential uncertainty quantification. Arxiv: 2307.14675 (2023)
7. Howland, M.F., Dabiri, J.O.: Wind farm modeling with interpretable physics-informed machine learning. Energies **12**(14) (2019)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
9. von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S.: Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering **60**, 86–101 (2014)