

Towards transparent and robust data-driven wind turbine power curve models

Simon Letzgus¹ and Klaus-Robert Müller^{1,2,3,4}

¹Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany

²BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

³Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea

⁴Max Planck Institute for Informatics, Stuhlsatzenhausweg 4, 66123 Saarbrücken, Germany

Abstract—Wind turbine power curve models translate ambient conditions into turbine power output. They are essential for energy yield prediction and turbine performance monitoring. In recent years, data-driven machine learning methods have outperformed parametric, physics-informed approaches. However, they are often criticised for being opaque “black boxes” which raises concerns regarding their robustness in non-stationary environments, such as faced by wind turbines. We, therefore, introduce an explainable artificial intelligence (XAI) framework to investigate and validate strategies learned by data-driven power curve models from operational SCADA data. It combines domain-specific considerations with Shapley Values and the latest findings from XAI for regression. Our results suggest, that learned strategies can be better indicators for model robustness than validation or test set errors. Moreover, we observe that highly complex, state-of-the-art ML models are prone to learn physically implausible strategies. Consequently, we compare several measures to ensure physically reasonable model behaviour. Lastly, we propose the utilization of XAI in the context of wind turbine performance monitoring, by disentangling environmental and technical effects that cause deviations from an expected turbine output. We hope, our work can guide domain experts towards training and selecting more transparent and robust data-driven wind turbine power curve models.

Index Terms—Explainable AI (XAI), Machine Learning, Wind Energy, Wind Turbine Power Curve, SCADA, Condition Monitoring

I. INTRODUCTION

The energy sector is responsible for the majority of global greenhouse gas emissions [46] and wind energy is to play a key role in its decarbonisation. The globally installed capacity has surpassed the 800 GW mark and is expected to double over the next decade [13]. Accurate wind turbine power curve models are crucial enablers for this transition. Coupled with meteorological forecasts, they are key for short- and long term energy yield predictions which are essential for grid operation and planning, energy trading, and investment decisions [36], [35]. Moreover, power curve models can be utilized for wind turbine condition- and performance monitoring [24], [51].

Therefore, power curve modelling has received significant attention [57], [14]. Early approaches have mainly focused on parametric models that follow physical considerations (e.g. [44], [15], [24], [56]). More recently, they were outperformed by data-driven machine learning (ML) models which have become the state-of-the-art [32], [51], [43], [36], [35], [38],

[42], [6]. It is, however, difficult to convey the implicit strategy of such complex, non-linear ML models to the user [49], [17]. Additionally, they usually lack explicit causal or physical understanding of the data-generating process and can capture any pattern in the training data that improves performance [25], [2]. Naturally, this raises concerns regarding the models’ ability to generalize beyond the data seen during training, especially in highly non-stationary environments like the wind energy domain [27]. Therefore, the wind community has often uttered the need for more transparency of data-driven approaches [60], [57], [36], [9], [6].

At the same time, eXplainable AI (XAI) has become a major subfield of ML (see [50], [49], [33] for reviews), enabling the user to gain an understanding of the decision process of some of the most complex ML models. Recently, these methods have started to be applied in the wind energy domain as well [10]. For example, in the context of wind turbine monitoring [9], [55], [34] and power prediction [61], [39]. However, to our knowledge, no prior work so far has utilized XAI methods to systematically analyse, compare and validate strategies learned by data-driven power curve models from operational wind turbine data. We address this gap and demonstrate how to use model explanations in practical applications. These include training and selecting more robust data-driven power curve models, deciding whether we have collected enough training data, evaluating our data pre-processing pipeline and explaining deviations from an expected turbine power output.

First, we review the relationships between environmental parameters and a wind turbine’s power output, as well as existing power curve modelling approaches in section II. Then, we describe how to calculate domain-specific Shapley values for intuitive and quantitatively faithful power curve model explanations (see Sec. III). After introducing the data set, models, and experimental settings (see Sec. IV), we compare strategies applied by physics-informed and data-driven models (see Sec. V). There, we also draw connections between model strategies and their generalization ability (see Sec. V-B. Afterwards, we investigate different approaches for obtaining more physically plausible data-driven power curve models (see Sec. VI). Section VII adds an application case of Shapley values in the context of turbine performance monitoring. Finally, Section VIII concludes the paper with a concise summary and discussion.

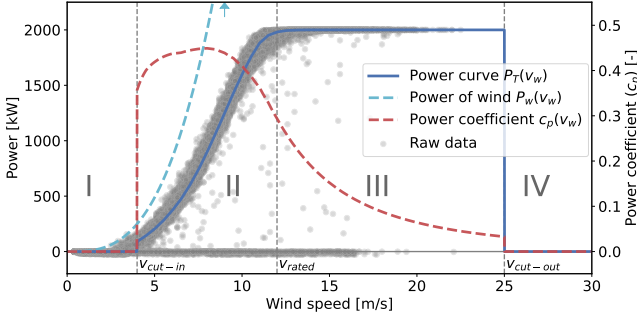


Fig. 1. Power of wind (light blue, dashed) vs. wind speed and a turbine's power curve under standard conditions (dark blue), from which the power coefficient (c_p) can be derived. Measured data points are also displayed (grey markers). Additionally, the four distinct operational regions are marked.

II. UNDERSTANDING AND MODELLING POWER CURVES

In this section, we review the physical basics of wind energy conversion and power curve modelling to facilitate the interpretation of model strategies later on.

A. From wind to power

The kinetic power of the wind (P_w) can be derived using the formulas of classical mechanics [16]. It depends on air density (ρ), the area swept by a turbine's rotor (A_r) and, of course, the wind speed (v_w):

$$P_T = \underbrace{\frac{1}{2} \rho A_r v_w^3}_{\text{Power of wind } (P_w)} c_p(\lambda, \beta). \quad (1)$$

Wind turbines are designed and controlled to extract a maximum of the available wind power and convert it to electricity. The share of power extracted by a wind turbine's rotor is described by the power coefficient c_p , which depends on aerodynamic properties of the rotor, such as the tip-speed-ratio λ and blade pitch angle β . Figure 1 shows that in practice, c_p is zero below a certain minimum (cut-in) wind speed (operational region I), where the turbine cannot extract enough energy to overcome its initial rotor momentum. It reaches a maximum between cut-in and rated wind speed (operational region II). At rated wind speed, the blades are deliberately pitched away from the aerodynamic optimum not to exceed the nominal power of the generator (operational region III). Finally, turbines are shut down for safety reasons at very high (cut-out) wind speeds (typically >25 m/s, operational region IV).

Together, this results in the typical shape of a wind turbine's power curve, which describes its nominal power output for given environmental conditions. It is usually displayed in a power versus wind speed plot (Fig. 1), assuming fixed values for all the other environmental factors.

B. Measuring power curves

IEC 61400 12-1 [19], a widely adopted international standard, describes in detail how to measure power curves in practice. The basis represents the so-called binning method.

Wind speed measured at hub height in 10min intervals and average production is calculated over 0.5 m/s wind speed bins. It also specifies further details regarding measurement and calculation as well as several correction methods, some of which we introduce in the following, to account for the simplified assumptions taken by the binning method.

1) *Air density correction*: To account for the influence of air density, a straightforward correction method can be derived from Equation 1. The measured 10min wind speed (v_w) is normalized by applying Equation 2 where ρ_t is the air density at the respective point in time and ρ_{mean} the mean air density across the full measurement period. Air density itself can be calculated using ambient temperature, air pressure and relative humidity [19].

$$v_{w,\rho} = v_w \left(\frac{\rho_t}{\rho_{mean}} \right)^{1/3} \quad (2)$$

2) *Turbulence correction*: Note that Equation 1 implies a non-trivial simplification: the highly complex atmospheric turbulent flow in the rotor plane is characterized by a single equivalent wind speed [12]. Temporal wind speed variations over the 10min measurement intervals are typically characterized by turbulence intensity (TI) (Eq. 3). Their effect on the power curve depends on the operating region. Well below v_{rated} , higher turbulence increases turbine output, while the opposite effect is observed for wind speeds close to v_{rated} [1], [22]. IEC 61400 describes in detail a procedure to normalize a power curve to a given TI level ([19], Appendix M). It includes the iterative search for a zero turbulence power curve which enables the simulation of the turbine's output $P_{sim,TI}(v)$ for a given wind speed distribution and, therefore, TI level. Consequently, every bin-based power output $P(v)_{bin}$ can be normalized by a correction term which corresponds to the difference between the simulated turbine's output at the measured TI ($P_{sim,TI}(v)$) and the reference TI -level ($P_{sim,TI_{ref}}(v)$), respectively (Eq. 4). Despite being part of the standard since its update in 2017, the adoption of incorporating TI normalization has not become the norm in the wind industry so far [27].

$$TI = \frac{std(v_{w,10min})}{mean(v_{w,10min})} \quad (3)$$

$$P_{TI}(v_w) = P(v_w)_{bin} - \underbrace{P_{sim,TI}(v_w) + P_{sim,TI_{ref}}(v_w)}_{\text{TI correction term}} \quad (4)$$

C. Data driven modelling of power curves

Modelling wind turbine power curves is an active field of research (for reviews, see [57], [14]). Early approaches have mainly applied parametric models, such as logistic functions, which naturally resemble the shape of a power curve, with wind speed as the only model input [44], [15]. More recent contributions have included additional environmental input variables, which has turned the relatively simple uni-variate curve-fitting- into a multivariate regression problem [26], [36], [11], and applied many of the ML workhorses, such as

TABLE I
SELECTED ANN CONFIGURATIONS FOR MULTIDIMENSIONAL POWER CURVE MODELLING

	# Inputs	# Layers	# Parameters	Act. Functions	Notes
[29] (2001)	4	1	49	hyperbolic	Scaling of inputs due to phys. considerations
[51] (2013)*	3	2	28	sigmoid	*Architecture utilized in this work.
[43] (2016)	6	6	133	hyperbolic	multi-stage training due to phys. considerations
[36] (2019)*	7	3	13.451	relu	*Architecture utilized in this work.
[35] (2020)	5	4	6.891	relu	-

Random Forests (RF), Gaussian Processes, Support Vector Machines, and a variety of different Artificial Neural Network (ANN) architectures [56], [43], [38], [37], [42], [6], the latter being a popular and often best-performing choice [29], [31], [57], [43], [36], [35]. This has resulted in systematic improvements of power curve models, which benefit both predominant applications, power prediction [36], [35] and performance monitoring [24], [51]. Table I gives an overview of selected publications using ANN models and their respective model configurations which exemplify the trend towards more input variables and increasingly complex models.

The challenge of limited transparency of data-driven power curve models has been highlighted by several authors and different strategies for its mitigation have been developed in the past. For example, data pre-processing [29], simulated training data [11], multi-stage training procedures [43], the application of partially interpretable models [36] and some initial attempts to apply methods from the XAI domain. In [39], for example, SHAP is applied to assess the impact of different environmental parameters qualitatively and in [55], to calculate an indicator for long-term performance degradation. We extend these efforts beyond the qualitative analysis by incorporating the latest findings of XAI for regression problems [28] and conducting a systematic comparison between physical and data-driven model strategies.

III. EXPLAINING POWER CURVE MODELS

In this section, we introduce Shapley values and discuss their application to power curve models including the evaluation of learned model strategies.

A. Shapley values

Shapley values represent an intuitive and axiomatic approach to explain complex models [54], [58], [30]. The framework originated in game theory, where the related problem of sharing the total gain between a set of cooperating players was considered. The approach determines the contribution of a player (or feature) by removing it and observing the averaged difference with and without for all permutations:

$$A_i = \sum_{S|i \notin S} \alpha_S \cdot [f(x_{S \cup \{i\}}) - f(x_S)]$$

where x is a data point composed of N features. $\sum_{S|i \notin S}$ is a sum over all subsets of features that do not contain feature i , and x_S the data point x where only features in S have been retained (the other features have been set to zero or the value

of some meaningful reference point \tilde{x}). The normalisation constant $\alpha_S = |\mathcal{S}|!(N - |\mathcal{S}| - 1)!/N!$ ensures conservative explanations, meaning that $\sum A_i = f(x)$. The Shapley value can be applied to any function $f(x)$, whether it is a neural network, a random forest, or a physical model.

B. Meaningful Shapley values for power curve models

Recent research on XAI highlights two fundamental issues to be considered when explaining regression models [28]: The desirable completeness property, which enables the association of attributions with a physical unit (given for Shapley values) and the choice of a suitable reference-point \tilde{x} (and corresponding reference value $\tilde{y} = f(\tilde{x})$) relative to which we explain. The latter is an active subject of research [59], [21] but unfortunately has not been addressed by any of the contributions that use Shapley values in the wind domain (compare [9], [55], [34], [61], [39]). There, the standard choice of the popular SHAP package [30] which is the mean input feature vector \bar{X}_{tr} over all samples is utilized as some kind of a "neutral" baseline in the absence of any further information. However, this out-of-the-box application of the method limits the insights to qualitative importance ranking of features and hinders a detailed evaluation of learned model strategies. Therefore, we advocate domain-specific settings for the application to wind turbine power curve models which depend on the concrete application case.

To validate global model behaviour, we suggest explaining relative to the minimum feature vector $\min(X_{tr})$. This generates intuitive explanations relative to wind speed zero (or cut-in, depending on data pre-processing) and therefore relative to zero-power output in absolute terms. Moreover, it facilitates the validation of attributions based on their sign, depending on the different operational regions (Fig. 1). Furthermore, a problem-specific, conditioned choice of \tilde{x} , similar to [59], enables to explain deviations from an expected power output (compare Sec. VII). This can help, for example, to understand why data-driven models perform better than physical models or to meaningfully attribute turbine underperformance to features.

C. Evaluating XAI attributions for power curve models

A remaining challenge and active field of research is, how to evaluate and present XAI attributions to domain experts for maximal benefit [17], [49]. When evaluating the physical compliance of data-driven model strategies, we leverage the model-agnostic nature of Shapley values. We generate attributions for both, physical and data-driven models, and

compute correlation coefficients between them (R_{phys}^2). Using similarity to physical models as a performance metric has one obvious shortcoming: any deviation from the physical model, even if it addresses a known weakness, results in a decrease. Therefore, it is not the aim to maximize correlation but rather to avoid poor correlation which indicates a clear violation of the known physical principles. In this sense, the correlation has shown to be a suitable and meaningful quantitative metric. Moreover, we evaluate the quantitative faithfulness of single attributions when explaining deviations from an expected turbine output (see Sec. VII). For that, we augment the data in a physics-informed manner. This controlled environment then allows for a comparison of magnitudes between attributions and ground truth.

Beyond quantitative evaluation, suitable visualization techniques are required to facilitate the manual interpretation of attributions by domain experts. We propose to plot attribution distributions conditioned on the measured wind speed $P(A_i|v_w)$ against the wind speed. This results in global explanations with a high resemblance to the way power curves are typically displayed and therefore facilitates contextualisation and interpretation by domain experts (compare Fig. 2). For the analysis of single data points, simple bar plots are sufficient (compare Fig. 5).

IV. EXPERIMENTAL SETUP

In this section, we introduce the data sets, models and the experimental setup of the subsequent analysis.

A. Data, Preprocessing & Feature Selection

We use openly accessible operational data from the SCADA system of four 2 MW wind turbines¹ (called Turbine A to D, respectively ²) and a meteorological met-mast, all located within the same site. The sensors (typical 10min averaged resolution) and logs cover two full years of operation. Our preprocessing consists of several basic filters which remove periods where data was incomplete, non-operational periods based on a simple power production threshold of 0 MW, and data points affected by curtailment or stoppages based on respective SCADA-log-messages. Overall, around 50.000 data points per turbine remain (approximately 50% of the original size), which we then temporally divide into train and test sets (one full year each), as well as a validation set (20% randomly sampled from the training data).

As model inputs, we select **wind speed** (v_w), **air density** (ρ) and **turbulence intensity** (TI). This limits complexity and enables a fair comparison with the physical baseline model (both in terms of performance and strategy). Note, that we normalize the inputs with a min/max-scaling and explain relative to the reference point $\tilde{x} = 0$ when analyzing model strategies. Therefore, we yield attributions that sum up to the function output minus its bias. When explaining deviations from an expected output (Sec. VII), however, we explain relative to individually chosen reference points.

¹<https://opendata.edp.com>

²Turbine A to D correspond to T01, T06, T07, and T11 of the respective dataset. T09 was excluded due to its missing SCADA-log data.

B. Overview Models & Performance

Physical (IEC) model: we create a benchmark model based on physical considerations. For each turbine, we calculate the binned power curve (PC_{bin}) and apply air density as well as TI corrections following the IEC standard [19] ($PC_{\rho,TI}$, compare Sec. II-B). Before performing the TI correction, we apply a simple pruning and bias correction for the nacelle-measured TI values to match the TI distribution measured by the nearby met mast. All required parameterisation for the physical model (binned power curve, average air density and zero TI-reference power curve) are calculated using the training data set.

TABLE II
SUMMARY OF MODEL PERFORMANCE (TEST SET RMSE [kW])

Model	Turbine A	Turbine B	Turbine C	Turbine D
IEC_{bin}	49.65	41.53	48.67	51.07
$IEC_{\rho,TI}$	43.60	35.40	40.36	41.6
RF	36.67 ± 0.03	34.16 ± 0.02	34.52 ± 0.03	36.73 ± 0.04
ANN_{small}	35.67 ± 0.54	32.89 ± 0.37	$32.76.92 \pm 0.53$	36.46 ± 0.37
ANN_{large}	34.50 ± 0.08	32.73 ± 0.30	31.79 ± 0.16	35.52 ± 0.17

Random Forest (RF): RFs are a popular decision tree-based ensemble method. They have successfully been applied in power curve modelling (e.g. [24], [20], [38]). Therefore, we chose them as a well-established data-driven baseline. Each consists of 100 estimators and is regularized with a minimum of 30 samples for a split and 3 to form a leaf.

Artificial Neural Networks (ANN): Historically, small architectures with only a few neurons in up to two hidden layers have shown to be suitable candidates (compare e.g. [51], [31]). We include the best-performing ANN architecture from [51] as a representative of this model class (ANN_{small} : two layers with (3,3) neurons and sigmoid activations). Lately, larger fully-connected, feed-forward ANNs with multiple hidden layers and ReLU activations have become state-of-the-art in power curve modelling (see Sec. II-C). We, therefore, also include such a model into our comparison (the best-performing ANN architecture from [36] - ANN_{large} : three layers with (100,100,50) neurons and ReLU activations, $\lambda = 0.05$). We train ANNs with the Adam optimizer [23] (adaptive learning rate, early stopping after 100 epochs).

Data-driven models were implemented with `scikit-learn` [41] and hyperparameters which were not specified in the respective publications (training modalities, for example), were selected based on a grid-search with 5-fold cross-validation on the training data set. Table II facilitates a test-set performance comparison of the different models. As expected, the IEC corrections clearly improve the binned power curves, although there is considerable variance between turbines. T_A data seems to deviate the most from the physical model, T_B the least. In line with the literature, all data-driven models outperform the IEC models and large ANNs show consistently lowest root mean squared errors (RMSEs), followed by their small counterparts and the RFs. This ranking holds across all turbines.

TABLE III

CORRELATION OF STRATEGIES BETWEEN PHYSICAL AND DATA-DRIVEN MODELS, PER MODEL TYPE, TURBINE AND INPUT VARIABLE. MEAN AND STANDARD DEVIATIONS OVER 10 RANDOMLY INITIALIZED TRAINING RUNS. AVERAGE R^2_{phys} OVER ALL MODELS, TURBINES AND RUNS IS 0.77.

Model	Turbine A				Turbine B			
	$R^2_{v_w}$	R^2_{ρ}	R^2_{TI}	$R_{mean,max,min}$	$R^2_{v_w}$	R^2_{ρ}	R^2_{TI}	$R_{mean,max,min}$
RF	1.00 ± 0.0	0.70 ± 0.01	0.30 ± 0.01	$0.67^{0.66}_{0.67}$	1.0 ± 0.0	0.91 ± 0.00	0.63 ± 0.00	$0.84^{0.84}_{0.85}$
ANN_{small}	1.00 ± 0.0	0.79 ± 0.11	0.40 ± 0.26	$0.73^{0.86}_{0.51}$	1.0 ± 0.0	0.93 ± 0.02	0.83 ± 0.03	$0.92^{0.93}_{0.89}$
ANN_{large}	1.00 ± 0.0	0.53 ± 0.10	0.22 ± 0.03	$0.58^{0.63}_{0.54}$	1.0 ± 0.0	0.94 ± 0.01	0.67 ± 0.04	$0.87^{0.88}_{0.85}$

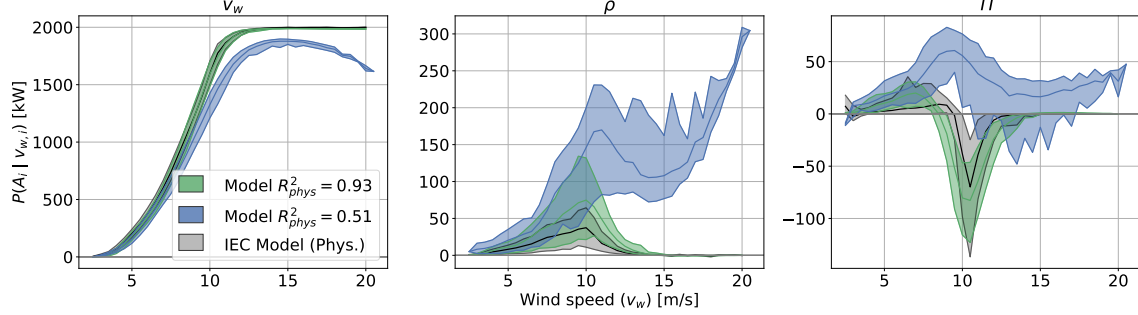


Fig. 2. Conditional distributions of attributions (mean as lines, range shaded) for the physical model (grey) and the data-driven models with the lowest (blue) and the highest (green) similarity to the physical strategy. The curves illustrate the wide range of adapted strategies and exemplify the findings presented in Table III on the level of individual models.

V. OPENING THE BLACK BOX: INSIGHTS INTO DATA-DRIVEN POWER CURVE MODELS

In this section, we systematically analyse data-driven power curve models using Shapley attributions to validate their internal strategy against a physical baseline (IEC model). We focus on a comparison between T_A and T_B which represent the extremes in terms of data alignment with the physical model (compare Table II). Moreover, we investigate the connection between model strategy and robustness.

A. Overview of data-driven model strategies

Table III shows the similarity of data-driven and physical model strategies for the different models, turbines and input parameters. We observe a wide range of correlation between data-driven and physical model attributions which ranges from 0.51 up to 0.93 (average across all models and both turbines is 0.77). To get a better qualitative understanding of what these correlations mean, we display the conditional distributions for both extremes (plus the IEC model as a reference) in Figure 2. We can observe, that a R^2_{phys} of 0.93 corresponds to capturing the physical relationships in almost a textbook manner (green) while the R^2_{phys} of 0.51 means failing in most aspects (blue). For more systematic insights, we look at the various impact factors separately:

Impact of turbine: the presented extreme cases are representative when it comes to the impact of the different turbines. Data-driven strategies agree consistently much less with the physical strategies on T_A than they do on T_B . This indicates that the data from T_A contains traces of effects the IEC model does not explicitly account for (IEC model performance

pointed in this direction, see Tab. II) which, sometimes, breaks physically reasonable behaviour in the textbook sense. Note, however, that some models are still able to learn strategies with an R^2_{phys} of up to 0.86.

Strategies by input feature: in general, we can observe a clear ranking in mean and variance of R^2_{phys} for the respective input parameters. The characteristic influence of v_w , the by far most crucial input, is captured very accurately across models and turbines. The influence of ρ is mostly captured reasonably well. Interestingly, we observed that ANNs mainly overestimate the influence of ρ compared to the physical baseline (a moderate and an extreme case can be seen in Fig. 2, centre). Whether they capture additional effects related to seasonality or this is caused only by spurious correlations remains up to speculation. For TI the data-driven models struggle the most and partially fail to account for its influence in a physics-informed manner, even though some follow the physical baseline very well (Fig. 2, right).

Impact of model-type: among the three different data-driven model types, the ANN_{small} shows the highest average R^2_{phys} on both turbines. However, it also produced the widest range of strategies for the different training runs, and therefore initializations (despite identical training data between runs). This effect is particularly pronounced on T_A and can be observed for both ANN architectures, even for training runs that do not vary significantly in terms of RMSE. RFs, as an ensemble method, are much more consistent in their adopted strategy. On average, they have higher R^2_{phys} than ANN_{large} but never clearly outperform them in terms of maximally observed R^2_{phys} .

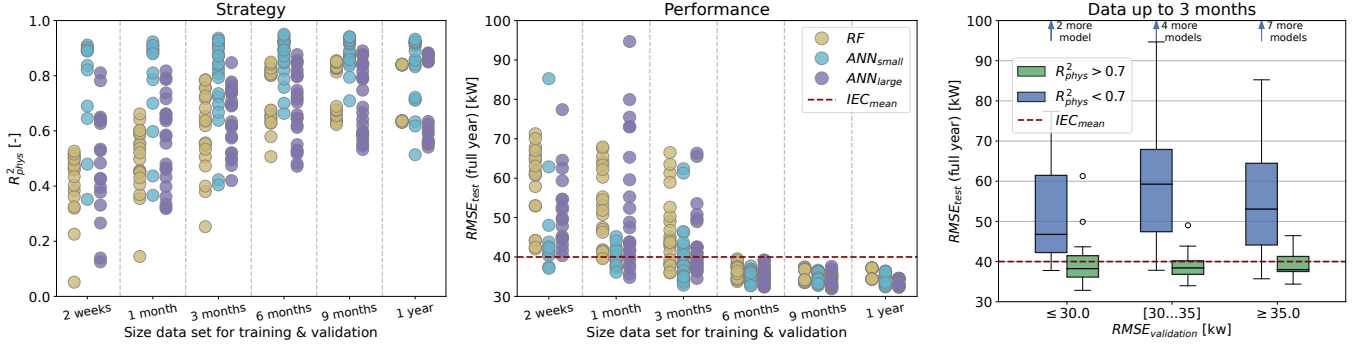


Fig. 3. Model strategy and performance beyond input distributions for both turbines (T_A & T_B). **Left:** R^2_{phys} of models trained and validated on the indicated period of data per model type. **Center:** Model performance on full-year test set when trained and validated on the indicated period. Note that models with only two weeks of training data outperformed the IEC model. **Right:** Full-year test set performance of all models trained and validated with up to three months of data. Validation performance in three bins on the x-axis. Colour indicates models with more (green) and less (blue) physically reasonable strategies.

B. The benefits of physically reasonable model strategies

Looking at Tables II and III, we realize that the model with the best test set performance (ANN_{large}) actually follows the least physical strategy on average. This naturally raises the question of what the benefits of physically reasonable strategies are. Originally, we were concerned about model generalization in a highly non-stationary environment. Therefore, we simulate this situation by conducting an ablation study where we artificially shorten the training period while still evaluate on the whole test year.

For each of the continuous training and validation periods $p_{tr,val} \in \{0.5, 1, 2, 3, 6, 9, 12\}$ months (the latter representing our original case from before), we train 12 models of each architecture. The training periods are spread equally over the training year and therefore partially overlap for periods longer than 1 month. 20% of each set was randomly selected and held back for validation. In Figure 3, left, we show the R^2_{phys} for the models that converged to training RMSEs of less than 100 kW (not all of them did) for both turbines. Interestingly, we observe that even with only two weeks of data we *can* train ML models that learn reasonable strategies, given sufficiently representative data (the ‘right’ two weeks) and a suitable model choice. With more training data, strategies get more consistent (partially due to the overlap of training data) and physically plausible on average. Moreover, two distributions of strategies emerge for the two turbines, which points towards a temporal concentration of the effects not described by the IEC model in the T_A data.

At the same time, Figure 3, centre shows that our concern regarding potentially high errors when going out of distribution is valid. Many of the models trained on periods less than 6 months do not generalize well. Only with training data of 6 months and more the data-driven models are consistently better than the IEC model (which still used the full year of data to calculate the binned PC). In this light, it is all the more surprising that some of the models trained with as little as two weeks of data were able to outperform the IEC model on the test set consisting of a full year.

Given these findings, we would like to identify these robust models effectively. Figure 3, right, shows the generalization

error for the subset of models trained and validated on periods of only up to three months. We distinguish between models with more (green) and less (blue) reasonable strategies and organize them based on their respective validation errors ($RMSE_{validation}$) into three bins. The results clearly show that it is indeed the models with physically reasonable strategies that generalize well beyond their own training distributions, which confirms our original hypotheses. At training time, however, we usually rely on cross-validation errors to test models for generalization in practice. The results demonstrate that this can be misleading in highly non-stationary environments. Low validation errors were not able to ensure robust models and model strategies were a much better indicator. These findings confirm that physically reasonable strategies are indeed desirable and call for a more prominent role of XAI in model selection to obtain more robust data-driven models.

VI. TOWARDS MORE PHYSICALLY PLAUSIBLE DATA-DRIVEN MODELS

In this section, we analyse different approaches to obtain more physically plausible data-driven models. We focus our analysis on T_A , where we observed the largest deviations from physical model strategies on average, and ANN_{large} , which performed best in terms of RMSE but worst in R^2_{phys} .

A. Regularization

Since the small ANN learned more physical model behaviour than the large ANN, a straightforward idea is to regularize the large ANN to obtain similar or even better results. We investigate the common L2-regularization and propose a physics-informed early stopping procedure.

L2-regularization: in Figure 4, left, we display R^2_{phys} and test set RMSE for different levels of regularization (λ). Interestingly, the larger the regularization the better the ANN captures the influence of air density at the cost of incorporating TI reasonably. The strategy for wind speed remains unchanged. This can have a beneficial effect on the overall model strategy which reaches a maximum at $\lambda = 0.25$ with an R^2_{phys} of 0.70 which is slightly below the native level of ANN_{small} at comparable test set performance.

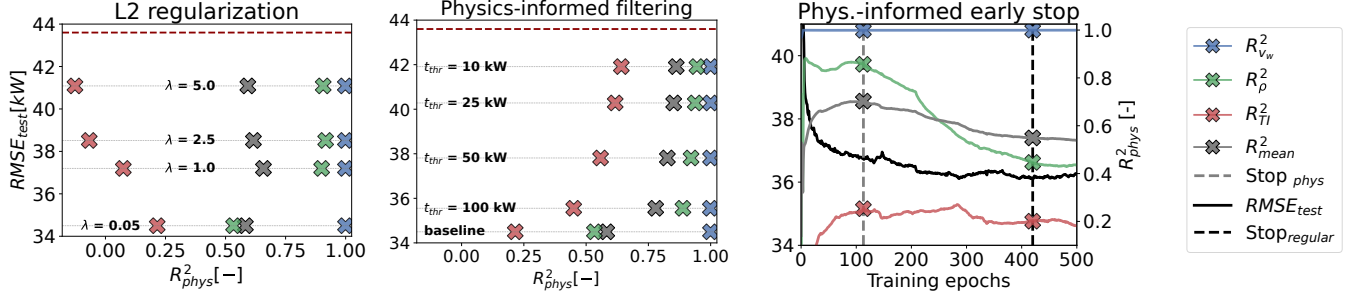


Fig. 4. Comparison of methods for physically more plausible models on T_A/ANN_{large} . **Left:** increasing L2 regularization beyond the baseline of $\lambda = 0.05$ initially yields a moderate increase of average R^2_{phys} , mainly driven by the more physical incorporation of ρ . For larger λ s this effect is overcompensated by the decrease in the model’s ability to physically reasonably account for TI . Markers represent the mean over 10 training runs. **Centre:** models trained on subsets of the training data, where $\hat{y}_{IEC} - y < t_{thr}$. The more narrow the filter, the more similar the ANN becomes to the physical model in both, strategy and performance. Markers represent the mean over 10 training runs. **Right:** traditional early stopping (black, dashed) aborted this training run much later than its physics-informed counterpart (grey, dashed), which also achieved significantly higher R^2_{mean} at only moderately increased $RMSE_{test}$.

Physics-informed early-stopping: analogously to traditional early stopping, which aborted training on average after around 500 epochs, we propose monitoring model strategies during training and stopping at the “physically most plausible” epoch. R^2_{phys} was increased to 0.68 and training was shortened by more than 250 epochs on average. The mean test set RMSE over ten initializations was clearly higher though. In some training runs, however, significant increases in R^2_{phys} were achieved at only moderate increase in $RMSE_{test}$ (see Fig. 4, right). These findings encourage the use of physics-informed early stopping for more plausible model strategies, also in combination with other measures.

B. Incorporating prior knowledge

Physics-informed data filtering: data filtering is an intuitive measure to ensure valid model behaviour and has been widely used in the context of ML models trained on wind turbine SCADA data [63], [57], [5], [4], [36]. We apply a simple, physics-informed approach by removing all data points from the training set where the IEC model error exceeds a certain threshold $t_{thrIEC} \in \{100, 50, 25, 10\}$. In Figure 4, right, we show the effect on the learned strategies and (unfiltered) test set RMSE. Physics-informed data pre-processing can indeed be used to bias a data-driven model strategy towards being more physically plausible. We note that the filtering enabled us to improve model strategy beyond the level of ANN_{small} ($R^2_{phys} = 0.78$) at a comparable test set performance.

Combination of physical and data-driven models: Lastly, we analyse the combination of the physical and the data-driven models, meaning we train the ANN_{large} using the adjusted target $\tilde{y} = y - y_{phys}$ and combine their output for the prediction. This strong, physics-informed prior is most effective in keeping overall model strategies closer to the original physical model ($R^2_{phys} = 0.89$) at test-set performance comparable to the other methods. Moreover, this approach is equivalent to the re-training method presented in [28] and thereby allows a direct investigation of what makes the respective data-driven model better than its physical counterpart, by simply analysing attributions of the model trained on \tilde{y} .

TABLE IV
COMPARISON OF APPROACHES FOR MORE PHYSICALLY PLAUSIBLE ML MODELS (R^2_{phys} FOR ANN_{large} ON T_A). ORIGINAL MODELS (FIRST TWO COLUMNS) FOR COMPARISON.

	Original		Regularization		Prior	
	ANN large	ANN small	L2-Reg.*	Phys.-stop	Data filter **	ML+ Phys.
R^2_{phys}	0.58	0.73	0.70	0.68	0.78	0.89
$RMSE_{test}$	34.50	35.67	35.61	37.2	35.56	35.61

* $\lambda=0.25$; ** $t_{thr}=100kW$

C. Implications for model selection

Table IV facilitates a comparison between the different presented approaches and the two ANN architectures as trained originally. All presented methods were able to significantly increase the physical plausibility of ANN_{large} . Overall, the incorporation of prior knowledge biased the models more strongly towards the physical model strategy than the less directed regularization approaches. In fact, the latter were not able to achieve results better than using moderately sized ANNs in the first place. Data filtering and the combination of data-driven and physical models, on the other hand, enabled models with much more reasonable, and therefore robust, strategies at similar test set performance. Moreover, they represent flexible tools that can be used to bias data-driven models arbitrarily close to the physical model strategies. In both cases, however, domain experts have to make design choices, such as the selection of filters and thresholds for the pre-processing pipeline, or the level of ML-model regularization in case of combining models. The presented XAI methodology represents a tool that can guide them towards more informed decisions in the model selection process. At comparable test set RMSE, the model with a physically more reasonable strategy is always preferred and models with particularly low R^2_{phys} should be discarded entirely. Eventually, an appropriate level of model robustness has to be selected based on the expert’s judgement regarding the representativeness of the available data.

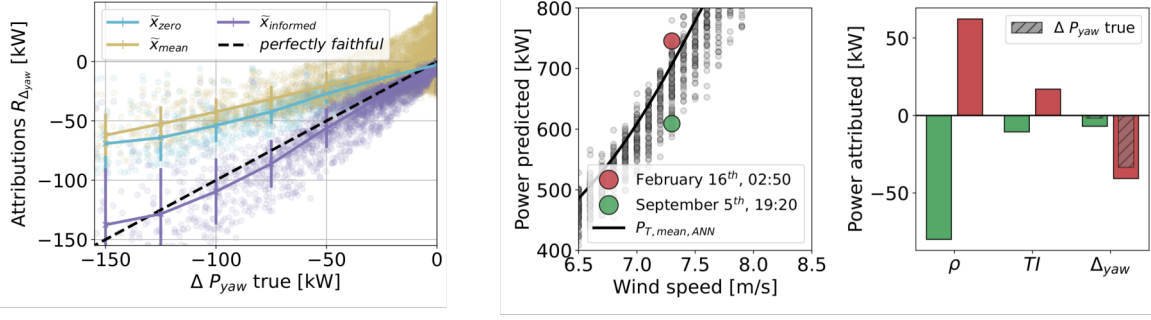


Fig. 5. **Left:** Faithfulness of attributions for different reference points (\tilde{x}_{ref}). The closer the attributions (points) to the diagonal dashed line, the more quantitatively faithful they are. The respective lines display mean and standard deviations over 25 kW bins. The informed reference point $\tilde{x}_{informed}$ clearly outperforms the others. **Center/Right:** Two data points are selected (centre) and the contributions of the different input features to the deviation from the learned power curve under mean conditions ($P_{T,mean,ANN}$) are displayed for both of them (right).

VII. EXPLAINING DEVIATIONS FROM AN EXPECTED TURBINE OUTPUT

In this section, we explain deviations from an expected turbine output, a highly relevant application in the context of performance monitoring [24], [52], [51], [8], [40]. Moreover, we demonstrate the importance of appropriate reference points for quantitatively faithful attributions.

We utilize a model of type ANN_{small} and include the absolute difference between average wind- and nacelle direction as a yaw misalignment feature (Δ_{yaw}). This allows for experiments in a controlled fashion by augmenting data with artificial yaw misalignment. We randomly add yaw misalignment of up to 15° to our data sets, and adjust the respective targets (turbine output) with a yaw misalignment factor $c_{ymis,i} = \cos^3(\Delta_{yaw})$, if $v_{w,i} < v_{rated}$. This approximation can be derived from equation 1, for more details on how yaw misalignment affects turbine output see [18]. After training and evaluation of the model on the augmented data, we can compare the magnitude of Shapley attributions to the ground truth:

$$\Delta P_{yaw,i}true = \begin{cases} c_{ymis,i} \cdot P_{T,i}, & \text{if } v_{w,i} < v_{rated} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

This is shown in Figure 5, left, where we explain the same ANN_{small} model using three different reference points \tilde{x}_{ref} . We can observe large differences in terms of absolute magnitudes in the resulting attributions. Both, the zero reference point which we used earlier for explaining global model strategies (\tilde{x}_{zeros}), as well as the mean vector over the training set (\tilde{x}_{mean}), often the standard choice, fail to attribute for the power reduction induced by yaw misalignment in a quantitatively faithful manner. This highlights the need to incorporate the assumptions implicit in the expected output, relative to which we explain. The informed reference point ($\tilde{x}_{informed}$) is conditioned on v_w : $x_{ref,i} = \mathbb{E}(x_i|v_w)$ for environmental parameters and set to zero for the yaw misalignment feature, which represents a healthy parameter baseline. This setting clearly outperforms both other reference points in terms of quantitative faithfulness.

Once we have ensured physically plausible model strategies and quantitatively faithful attributions, we can utilize explanations in a performance monitoring context (Fig. 5, centre and

right). The centre plot shows two selected data points along with the learned power curve under mean ambient conditions ($P_{T,mean,ANN}$ - no yaw misalignment). For the February instance (blue) the turbine produces around 40 kW more than the model standard power curve suggests, while during the September instance (green) it produces roughly 100 kW less, both at the same wind speed. Although both instances are towards the tails of the distribution, they are no clear outliers. But if we had to guess which of the two is affected by a technical malfunction, the September instance is a much more intuitive pick. The respective attributions (Fig. 5, right) reveal, however, that in February significant yaw misalignment was present but compensated by favourable ambient conditions. The September instance's lower output, on the other hand, can be attributed to less advantageous environmental conditions rather than a technical problem. The respective attributions enable a decomposition and quantification of different entangled effects (the instance attributions add up to the difference between $P_{T,mean,ANN}$ and $P_{T,i}$). This is particularly appealing in the context of turbine performance monitoring applications. Existing reconstruction-based anomaly detection approaches [48] (normal behaviour models, see e.g. [24], [52], [51], [8], [40]) have no such property. They would either include the yaw misalignment feature, which would cause them to "correctly" predict the under-performance and therefore not raise an alarm, or not include the yaw misalignment feature, correctly raise an alarm but without any indication of the potential root cause.

A trivial solution to the presented example would of course be to directly monitor the yaw misalignment error. However, this might not be possible for other potential influences on turbine power output that can be taken into account (think for example of blade-pitch-angles or turbine interactions). Moreover, the method is robust against poorly calibrated SCADA sensors. Nevertheless, the potential to indicate root-causes for under-performance is naturally limited to effects that are correctly captured by the model in the first place. Additionally, the absolute model error should serve as a confidence measure regarding the corresponding attributions (explanations for data points with low errors are more trustworthy).

VIII. SUMMARY & DISCUSSION

The flexibility of ML models is a curse and a blessing alike. In data-driven power curve modelling, it enables capturing turbine-individual behaviour and seamlessly incorporating additional input parameters, which results in impressive performance gains compared to parametric, physics-informed models. On the other hand, the models can capture any pattern in the training data that improves performance which, as we have shown, can be problematic in highly non-stationary environments such as those faced by wind turbines. In this contribution, we, therefore, introduced an XAI framework to validate strategies learned by data-driven wind turbine power curve models from operational SCADA data. We have found that models with physically plausible strategies are indeed more robust in out-of-distribution scenarios, explored measures to ensure physically reasonable strategies and demonstrated the value of explanations in the context of performance monitoring. This makes a strong case for utilizing XAI methods in data-driven wind turbine power curve modelling.

In practice, the presented methodology enables informed decisions along the model training and selection process. It can be used to monitor the effectiveness of data pre-processing measures, judge the quality as well as the amount of training data, and allows for the selection of models that are robust beyond the training distribution. Its value is exemplified by the demonstrated ability to select a model trained on only two weeks of data, which outperformed the physical model on a test set consisting of a whole year. Moreover, we found in our experiments that moderately sized ANNs (a few hidden layers with a few hidden neurons and sigmoid activations) learn on average more physically plausible strategies than deep, state-of-the-art ANNs and can therefore be considered a more robust model choice. This is particularly relevant in the light of the highly complex ML models proposed in the literature that were trained and evaluated on data significantly less than one year (e.g. [38], [36], [35], [37]). As we have seen, this may allow for models with small validation and test errors, but is at the risk of potentially poor generalization abilities and robustness.

As a ready solution, we have presented measures to successfully bias data-driven models towards more physically reasonable behaviour. Incorporating prior knowledge via data filtering or combining physical and data-driven models have shown to be particularly effective. They can be used to drive model strategies arbitrarily close to the physical model's (with the respective implications for test set performance, of course). But more generally speaking, with the presented methodology, we have introduced a novel tool that makes the effects of the respective design choices transparent and enables a more informed model selection by domain experts.

Moreover, we demonstrated the value of physically reasonable ML models in combination with quantitatively faithful attributions in the context of turbine performance monitoring. With the appropriate choice of ML model and reference point, attributions can decompose the deviation from an expected turbine output and assign it to the responsible input features. This can assist domain experts in the search for potential root causes of underperformance.

It is worth noting that the presented approach generally works for ML models of any complexity. For a significantly larger amount of inputs or lots of models, however, the calculation of exact Shapley values might become computationally too expensive. In this case, they can be replaced by an approximation [30] or computationally less expensive attribution methods, such as LRP [3] or PredDiff [7]. Also, choosing an appropriate physical model to have a baseline for every input can become more challenging in this scenario.

Future research could focus on using more elaborate physical benchmark models and validating the presented findings on a larger database and longer time horizons. Additionally, it would be interesting to see if regularization with an explanation-based penalty term, such as proposed by [47], [45], [53], can result in a better trade-off between model strategy and test set RMSE. Finally, we see a big potential for XAI attribution methods in the wind energy domain in general. They could, for example, contribute to further analysing and understanding turbine behaviour and interaction when being applied in the context of fluid dynamics and wake-modelling [62] or help to understand the gaps between simulated and measured data in many areas of wind research.

REFERENCES

- [1] A. Albers. Turbulence and shear normalisation of wind turbine power curve. *development*, 3:4, 1994.
- [2] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022.
- [3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 07 2015.
- [4] L. Bai, E. Crisostomi, M. Raugi, and M. Tucci. Wind turbine power curve estimation based on earth mover distance and artificial neural networks. *IET Renewable Power Generation*, 13(15):2939–2946, 2019.
- [5] P. Bangalore, S. Letzgus, D. Karlsson, and M. Patriksson. An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. *Wind Energy*, 20(8):1421–1438, 2017.
- [6] G. A. Barreto, I. S. Brasil, and L. G. M. Souza. Revisiting the modeling of wind turbine power curves using neural networks and fuzzy models: an application-oriented evaluation. *Energy Systems*, pages 1–28, 2021.
- [7] S. Blücher, J. Vielhaben, and N. Strodtloff. PredDiff: Explanations and interactions from conditional expectations. *Artificial Intelligence*, 312:103774, 2022.
- [8] S. Butler, J. Ringwood, and F. O'Connor. Exploiting scada system data for wind turbine performance monitoring. In *2013 Conference on Control and Fault-Tolerant Systems (SysTol)*, pages 389–394, 2013.
- [9] J. Chatterjee and N. Dethlefs. Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines. *Wind Energy*, 23(8):1693–1710, 2020.
- [10] J. Chatterjee and N. Dethlefs. Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future. *Renewable and Sustainable Energy Reviews*, 144:111051, 2021.
- [11] A. Clifton, L. Kilcher, J. K. Lundquist, and P. Fleming. Using machine learning to predict wind turbine power output. *Environmental Research Letters*, 8(2):024009, apr 2013.
- [12] A. Clifton, S. Schreck, G. Scott, N. Kelley, and J. K. Lundquist. Turbine inflow characterization at the national wind technology center. *Journal of solar energy engineering*, 135(3), 2013.
- [13] G. W. E. Council. Gwec—global wind report 2022. *Global Wind Energy Council: Brussels, Belgium*, 2022.
- [14] Y. Ding, S. Barber, and F. Hammer. Data-driven wind turbine performance assessment and quantification using scada data and field measurements. *Frontiers in Energy Research*, 10, 2022.

- [15] P. Giorsetto and K. F. Utsurogi. Development of a new procedure for reliability modeling of wind turbine generators. *IEEE Transactions on Power Apparatus and Systems*, PAS-102(1):134–143, 1983.
- [16] E. Hau. *Wind turbines: fundamentals, technologies, application, economics*. Springer Science & Business Media, 2013.
- [17] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.
- [18] M. F. Howland, C. M. González, J. J. P. Martínez, J. B. Quesada, F. P. Larranaga, N. K. Yadav, J. S. Chawla, and J. O. Dabiri. Influence of atmospheric conditions on the power production of utility-scale wind turbines in yaw misalignment. *Journal of Renewable and Sustainable Energy*, 12(6):063307, 2020.
- [19] IEC. Wind energy generation systems – part 12-1: Power performance measurements of electricity producing wind turbines. Standard IEC 61400-12-1, International Electrotechnical Commission, Geneva, Switzerland, 2017.
- [20] O. Janssens, N. Noppe, C. Devriendt, R. Van de Walle, and S. Van Hoecke. Data-driven multivariate power curve modeling of offshore wind turbines. *Engineering Applications of Artificial Intelligence*, 55:331–338, 2016.
- [21] D. Janzing, L. Minorics, and P. Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.
- [22] K. Kaiser, W. Langreder, H. Hohlen, and J. Højstrup. Turbulence correction for power curves. In *Wind Energy*, pages 159–162. Springer, 2007.
- [23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [24] A. Kusiak, H. Zheng, and Z. Song. On-line monitoring of power curves. *Renewable Energy*, 34(6):1487–1493, 2009.
- [25] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019.
- [26] G. Lee, Y. Ding, M. G. Genton, and L. Xie. Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *Journal of the American Statistical Association*, 110(509):56–67, 2015.
- [27] J. C. Lee, P. Stuart, A. Clifton, M. J. Fields, J. Perr-Sauer, L. Williams, L. Cameron, T. Geer, and P. Housley. The power curve working group’s assessment of wind turbine power performance prediction methods. *Wind Energy Science*, 5(1):199–223, 2020.
- [28] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon. Toward explainable artificial intelligence for regression models: A methodological perspective. *IEEE Signal Processing Magazine*, 39(4):40–58, 2022.
- [29] S. Li, D. C. Wunsch, E. A. O’Hair, and M. G. Giesselmann. Using neural networks to estimate wind turbine power generation. *IEEE Transactions on energy conversion*, 16(3):276–282, 2001.
- [30] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.
- [31] M. Lydia, A. I. Selvakumar, S. S. Kumar, and G. E. P. Kumar. Advanced algorithms for wind turbine power curve modeling. *IEEE Transactions on Sustainable Energy*, 4(3):827–835, 2013.
- [32] K. Methaprayoon, C. Yingvivatanapong, W.-J. Lee, and J. R. Liao. An integration of ann wind power estimation into unit commitment considering the forecasting uncertainty. *IEEE Transactions on Industry Applications*, 43(6):1441–1448, 2007.
- [33] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.
- [34] A. Movsessian, D. G. Cava, and D. Tcherniak. Interpretable machine learning in damage detection using shapley additive explanations. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 8(2):021101, 2022.
- [35] J. Nielson, K. Bhaganagar, R. Meka, and A. Alaeddini. Using atmospheric inputs for artificial neural networks to improve wind turbine power prediction. *Energy*, 190:116273, 2020.
- [36] M. Optis and J. Perr-Sauer. The importance of atmospheric turbulence and stability in machine-learning models of wind farm power production. *Renewable and Sustainable Energy Reviews*, 112:27–41, 2019.
- [37] R. Pandit, D. Infield, and M. Peñas. Accounting for environmental conditions in data-driven wind turbine power models. *IEEE Transactions on Sustainable Energy*, pages 1–10, 2022.
- [38] R. K. Pandit, D. Infield, and A. Kolios. Comparison of advanced non-parametric models for wind turbine power curves. *IET Renewable Power Generation*, 13(9):1503–1510, 2019.
- [39] C. Pang, J. Yu, and Y. Liu. Correlation analysis of factors affecting wind power based on machine learning and shapley value. *IET Energy Systems Integration*, 3(3):227–237, 2021.
- [40] J.-Y. Park, J.-K. Lee, K.-Y. Oh, and J.-S. Lee. Development of a novel power curve monitoring method for wind turbines and its field tests. *IEEE Transactions on Energy Conversion*, 29(1):119–128, 2014.
- [41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [42] S. Pei and Y. Li. Wind turbine power curve modeling with a hybrid machine learning technique. *Applied Sciences*, 9(22):4930, 2019.
- [43] F. Pelletier, C. Masson, and A. Tahan. Wind turbine power curve modelling using artificial neural network. *Renewable Energy*, 89:207–214, 2016.
- [44] W. R. Powell. An analytical expression for the average output power of a wind machine. *Solar Energy*, 26(1):77–80, 1981.
- [45] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020.
- [46] H. Ritchie, M. Roser, and P. Rosado. *co₂ and greenhouse gas emissions. Our World in Data*, 2020. <https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>.
- [47] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.
- [48] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.
- [49] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.
- [50] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019.
- [51] M. Schlechtingen, I. F. Santos, and S. Achiche. Using data-mining approaches for wind turbine power curve monitoring: A comparative study. *IEEE Transactions on Sustainable Energy*, 4(3):671–679, 2013.
- [52] M. Schlechtingen, I. F. Santos, and S. Achiche. Wind turbine condition monitoring based on scada data using normal behavior models. part 1: System description. *Applied Soft Computing*, 13(1):259–270, 2013.
- [53] X. Shao, A. Skryagin, W. Stammer, P. Schramowski, and K. Kersting. Right for better reasons: Training differentiable models by constraining their influence function. In *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [54] L. S. Shapley. *A Value for n-Person Games*, pages 307–318. Princeton University Press, 1953.
- [55] H. Shin, M. Rüttgers, and S. Lee. Neural networks for improving wind power efficiency: A review. *Fluids*, 7(12), 2022.
- [56] S. Shokrzadeh, M. J. Jozani, and E. Bibeau. Wind turbine power curve modeling using advanced parametric and nonparametric methods. *IEEE Transactions on Sustainable Energy*, 5(4):1262–1269, 2014.
- [57] V. Sohoni, S. Gupta, and R. Nema. A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems. *Journal of Energy*, 2016, 2016.
- [58] E. Strumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, 2010.
- [59] M. Sundararajan and A. Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.
- [60] J. Tautz-Weinert and S. J. Watson. Using scada data for wind turbine condition monitoring—a review. *IET Renewable Power Generation*, 11(4):382–394, 2016.
- [61] U. S. Tenfjord and T. V. Strand. The value of interpretable machine learning in wind power prediction: an empirical study using shapley additive explanations to interpret a complex wind power prediction model. Master’s thesis, Norwegian School of Economics, 2020.
- [62] Z. Ti, X. W. Deng, and H. Yang. Wake modeling of wind turbines using machine learning. *Applied Energy*, 257:114025, 2020.
- [63] L. Zheng, W. Hu, and Y. Min. Raw wind data preprocessing: A data-mining approach. *IEEE Transactions on Sustainable Energy*, 6(1):11–19, 2015.