Energy and AI 24 (2026) 100695

Contents lists available at ScienceDirect

Energy and AI

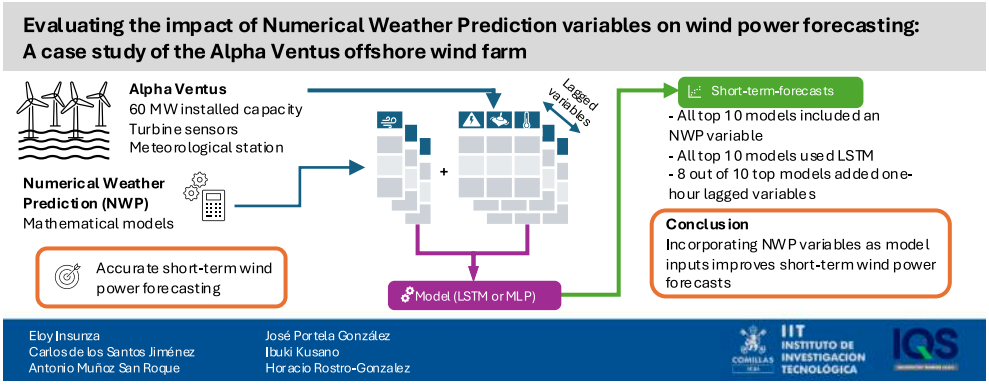
journal homepage: www.elsevier.com/locate/egyai

Evaluating the impact of Numerical Weather Prediction variables on wind power forecasting: A case study of the Alpha Ventus offshore wind farm

Eloy Insunza [a,*], Carlos de los Santos Jiménez [a], Antonio Muñoz [a], José Portela [a], Ibuki Kusano [b], Horacio Rostro-Gonzalez [b,c]

[a] Institute for Research in Technology (IIT), Technical School of Engineering (ICAI), Comillas Pontifical University, Madrid, 28008, Comunidad de Madrid, Spain
[b] GEPI Research Group, IQS-School of Engineering, Ramon Llull University, Barcelona, 08017, Cataluña, Spain
[c] Department of Electronics Engineering, DICIS–University of Guanajuato, Carretera Salamanca–Valle de Santiago km 3.5 + 1.8 kms, Salamanca 36885, Mexico

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Combining NWP and AI methods improved short-term offshore wind power forecasts.
- Both Multilayer Perceptron and Long Short-Term Memory architectures were evaluated.
- Long Short-Term Memory models consistently outperformed the Multilayer Perceptron.
- The analysis used real operational data from the Alpha Ventus offshore wind farm.
- Open-source datasets and machine learning were used to impute missing sensor data.

ARTICLE INFO

Keywords:
Numerical weather prediction
Offshore wind farms
Wind power forecasting
Long short-term memory
Deep learning

ABSTRACT

Offshore wind power generation has emerged as a reliable and stable source of renewable energy. However, accurate short-term forecasting of power generation remains a challenge due to the stochastic nature of weather conditions. This study evaluates the contribution of Numerical Weather Prediction (NWP) outputs and lagged explanatory variables for short-term offshore wind power forecasting. The analysis was conducted using data from the Alpha Ventus wind farm, located in the North Sea. NWP outputs from the ICON-D2 (ICOsahedral Nonhydrostatic D2) model were integrated with historical power generation data collected from Alpha Ventus

* Corresponding author.
E-mail address: einsunza@comillas.edu (E. Insunza).

https://doi.org/10.1016/j.egyai.2026.100695
Received 24 June 2025; Received in revised form 9 January 2026; Accepted 7 February 2026
Available online 14 February 2026

turbine sensors. The performance was evaluated using a state-of-the-art recurrent neural network (Long Short-Term Memory, LSTM) alongside four established machine learning baselines (Multi-Layer Perceptron, XGBoost, Random Forest, and LightGBM).

The results highlight three main findings. First, the inclusion of NWP predictors consistently improves performance across all evaluated technologies. Second, LSTM-based models improve forecasting accuracy compared to the alternative algorithms. Third, while adding 1 h lagged variables is beneficial, extending the lag structure beyond this does not yield additional gains in predictive performance. These findings emphasize the potential of advanced neural network architectures combined with NWP data to improve offshore wind power generation forecasting accuracy.

1. Introduction

The global energy landscape is undergoing a rapid transformation as the demand for sustainable and renewable energy sources becomes increasingly urgent. Among these, offshore wind power generation has emerged as a critical contributor due to its ability to harness the consistent and powerful wind currents found at sea. Unlike onshore wind farms, offshore installations benefit from minimal physical obstructions, leading to higher and more stable wind speeds [1]. This characteristic translates into a more reliable energy output, positioning the offshore wind as a key player in achieving global energy transition goals.

Despite these advantages, offshore wind farms face unique challenges. Foley et al. noted that offshore wind farms are highly dependent and sensitive to weather and atmospheric phenomena, significantly impacting their performance and operational costs [2]. Maintenance and longevity are particularly affected by these conditions, with the deployment and maintenance costs of offshore installations being approximately 1.5 to 2 times higher than those of onshore wind farms [3]. Additionally, offshore wind farms pose challenges in data collection and integration of generated energy into the grid, further complicating their operation [4].

One of the most pressing challenges in the sector is accurate short-term forecasting of power generation. Wind variability, driven by complex and dynamic weather systems, complicates the prediction of power output, making it challenging for grid operators to manage supply and demand effectively [5,6]. Short-term forecasting is critical to integrating wind power into the energy grid, reducing the reliance on additional power sources, and ensuring the stability and efficiency of power distribution networks [7]. In addition, accurate forecasts allow for better planning of maintenance operations and better economic dispatch, enhancing the overall viability of wind power projects [8].

Machine learning (ML) has emerged as a promising solution to address the stochastic nature of wind power forecasting [9,10]. Advanced ML algorithms can capture complex, non linear patterns in weather data that traditional statistical models often fail to detect [11]. In this regard, this study investigates the integration of numerical weather prediction (NWP) outputs with advanced neural network architectures to enhance the accuracy of short-term wind power prediction.

The main contributions of this work are threefold and directly relate to both power system operations and applied artificial intelligence. First, we quantify the added value of operational Numerical Weather Prediction (NWP) outputs for short-term offshore wind power forecasting using data from the Alpha Ventus wind farm and ICON-D2 model, and we assess improvements relative to identical models without NWP inputs. Second, we provide a systematic benchmark of a state-of-the-art recurrent model (LSTM) against four established machine learning alternatives (MLP, XGBoost, Random Forest, and LightGBM). LSTM is particularly effective at capturing temporal dependencies in time-series data, making it well-suited for forecasting applications such as wind power generation [12,13]. Third, we analyze the impact of incorporating lagged explanatory variables, showing that including a 1-h lag is beneficial, whereas extending the lag structure beyond 1 h does not yield additional gains in predictive performance.

By leveraging high-resolution weather forecasts from the ICON-D2 (ICOsahedral Nonhydrostatic D2) model and historical turbine sensor data from the Alpha Ventus offshore wind Germany farm located in the North Sea [14,15], the proposed approach aims to improve predictive accuracy under realistic operational conditions. This addresses a core energy-system requirement: reducing forecast uncertainty that propagates into unit commitment, balancing actions, reserve procurement, and market bids in systems with high wind penetration.

The rest of the paper is organized as follows: Section 2 provides a review of the application of machine learning techniques in the context of wind forecasting. Section 3 describes the forecasting models used in this research, including the theoretical framework and methodology applied in the experimental phase. Section 4 explains the data processing pipeline, detailing how the raw data was cleaned, reconstructed, and prepared for training and evaluation. Section 5 presents the results obtained from the experiments and includes a discussion that interprets the impact of different variables and model configurations. Section 6 employs SHAP values to examine how NWP variables impact the forecasted values. Section 7 compares the computational cost of the LSTM with the baseline algorithms. Finally, Section 8 summarizes the main conclusions and insights derived from the study, offering potential directions for future work.

2. Related works

Forecasting methods are typically categorized into numerical, statistical, Machine Learning (ML), and hybrid approaches. According to Tuncar et al. [16], forecasting needs vary depending on the time horizon, which can be classified into very short-term (0 – 30 min), short-term (30 min – 6 h), medium-term (6 h – 1 day), and long-term (1 day – 1 month), each serving specific operational and strategic objectives.

On the other hand, Numerical Weather Prediction (NWP) models utilize extensive meteorological data and sophisticated computational algorithms to simulate atmospheric states and forecast weather conditions [17]. By incorporating data from sources such as satellites, weather stations, and buoys, these models can accurately predict variables like wind speed and direction, which are essential for estimating wind energy. However, these methods are not ideal for shorter-term predictions due to their high dependency on precise initial conditions, long processing times, and relatively low spatial and temporal resolution [18,19]. These models, designed for larger time scales, may fail to capture the rapid and localized changes required for immediate predictions. Therefore, for very short-term forecasts, faster approaches such as statistical models or machine learning are preferred, as they can quickly adapt to instantaneous variations in atmospheric conditions [20].

Statistical models, based on time series analysis, use historical data to predict future outcomes. These models are particularly effective for short time horizons of minutes or hours, identifying random patterns in the data [21]. Techniques such as ARMA, ARIMA, and Box-Jenkins models are commonly employed. For example, Torres et al. [22] demonstrated that ARMA models could reduce errors in hourly wind speed predictions by 20% compared to persistence models. Other studies, such as Erdem et al. [23], explored ARMA-based methods, including component decomposition and vector auto-regression (VAR) models, which improve wind direction and speed predictions. While

simple and easy to implement, these models are highly sensitive to nonlinearity and nonstationarity in the data [24].

Machine Learning (ML) methods leverage meteorological data to predict wind generation. Neural networks, particularly Long Short-Term Memory (LSTM) networks, have gained prominence due to their ability to handle nonlinear relationships and time-series dependencies [25,26]. Su et al. [27] used LSTM for wind power prediction, achieving a 4.96% improvement in RMSE over traditional methods. Similarly, Yin et al. [28] compared multiple algorithms, finding that LSTM outperformed others, achieving an MAPE of 0.032% and RMSE of 0.0018. Hanifi et al. [29] applied LSTM to predict wind power at a Scottish turbine, outperforming ARIMA and persistence models by 3.93% and 5.1%, respectively. Ko et al. [30] achieved superior accuracy with a Bi-LSTM network for ERCOT data, surpassing ARIMA and other models in prediction accuracy.

Hybrid methods combine multiple approaches, often integrating machine learning with digital signal processing techniques. These methods decompose non-stationary data into stationary subseries for improved prediction accuracy. Zu et al. [31] used wavelet packet decomposition (WPD) to split wind power data into frequency subseries, training GRU networks for each band. This approach outperformed traditional GRU models for wind power prediction. Similarly, Su et al. [32] applied WPD to separate wind speed into high- and low-frequency components, training LSTM networks on low-frequency data and applying additional decomposition with ensemble empirical noise decomposition (EEMD) for high-frequency data. This approach improved prediction accuracy by including rotor speed as a variable. The work performed by Hanifi et al. [21], further extended this hybrid approach for offshore wind turbines. Using WPD, they decomposed wind speed and power data into low- and high-frequency components. Low-frequency components trained an LSTM network, while high-frequency components trained convolutional networks, achieving superior performance compared to other algorithms with the lowest MSE, RMSE, and MAE values.

3. Modeling approach

This section provides an overview of the methodologies and theoretical foundations employed in our experimental framework. We begin by introducing the Long Short-Term Memory (LSTM) network, a neural architecture specifically designed to capture temporal dependencies. Following this, we describe the evaluation metric used to assess model accuracy, the Mean Absolute Error (MAE). Finally, we present the hyperparameter optimization framework employed called Optuna.

3.1. LSTM

LSTM (Long Short-Term Memory) networks are a specialized form of recurrent neural networks (RNNs) designed to overcome the vanishing and exploding gradient problems commonly encountered in traditional RNNs. These issues hinder the learning of long-range dependencies in sequential data. LSTM networks were first introduced by Hochreiter and Schmidhuber in 1997 [33]. They are particularly effective in capturing long-term dependencies thanks to their unique architecture, which incorporates memory cells and gating mechanisms that regulate the flow of information.

This structure enables LSTMs to learn which information to retain or discard over long sequences, making them particularly powerful for tasks involving time series forecasting, natural language processing, and other sequence modeling problems. In our work, LSTM networks were employed to predict short-term wind power generation by capturing temporal patterns in the input data. Fig. 1 shows the schema of an LSTM cell.

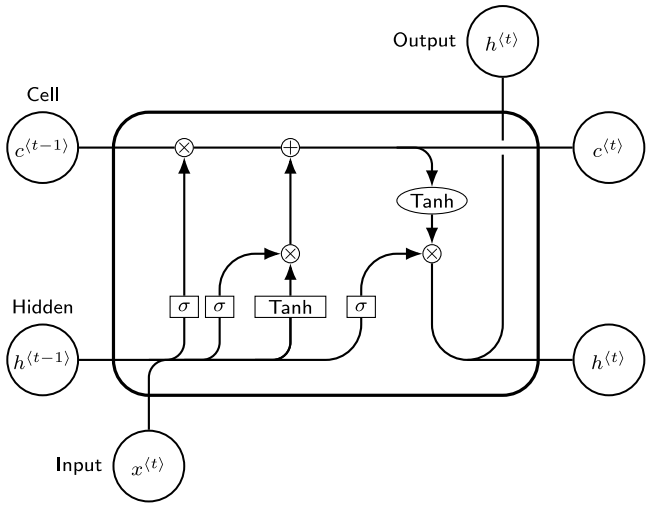


Fig. 1. Schema of an LSTM cell.

Table 1
LSTM and MLP hyperparameters and their ranges used for model optimization.

Hyperparameter	Range of values
Number of layers	[1, 2]
Neurons per layer	[1, 200]
Learning rate	[0.001, 0.1]
Batch size	{128, 256, 512}
Dropout	[0, 0.5]

3.2. Performance evaluation metric

We employed the Mean Absolute Error (MAE) as the performance indicator to evaluate the different forecasting models. Higher error values show the inaccuracy of a model. The corresponding equation is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

where:

- n is the total number of observations.
- y_i represents the true value for the ith observation.
- $\hat{y}_i$ represents the predicted value for the ith observation.

3.3. Hyperparameter optimization

Hyperparameter optimization [34–37] was conducted using the Optuna framework [38] to improve the machine learning models in this study. Optuna's define-by-run architecture dynamically creates the search space, allowing the process to adapt based on ongoing results. It uses two strategies to identify optimal hyperparameters. The sampling strategy focuses on promising areas of the hyperparameter space, utilizing probabilistic methods to enhance efficiency. The pruning strategy monitors the training process and stops unpromising trials early, saving computational resources. Table 1 details the hyperparameter ranges explored when training MLP and LSTM models. These ranges were chosen based on domain knowledge and initial experiments. Table 2 details the hyperparameter ranges explored when training the rest of the baseline models. The ranges for XGBoost and Random Forest are based on the work of Rajaperumal et al. [39], while the work of Bentéjac et al. [40] was used for LightGBM. The final optimized parameters were used to train models for robust predictions.

Table 2
XGBoost, Random Forest and LightGBM hyperparameters and the ranges used for their optimization.

XGBoost		Random Forest		LightGBM	
n_estimators	[100, 300]	n_estimators	[100, 300]	num_leaves	[3, 1024]
max_depth	[3, 9]	max_depth	[10, 30]	top_rate	[0.2, 0.7]
learning_rate	[0.01, 0.1]	min_samples_split	[2, 5]	learning_rate	[0.025, 0.3]
				other_rate	[0.05, 0.3]
				feature_frac tion_bynode	[0.25, 1.0]

Table 3
Comparison of Senvion 5M and Adwen AD-5116 wind turbines.

Attribute	Senvion 5M	Adwen AD-5116
Rated Power (MW)	5.075	5
Foundation	Jacket	Tripod
Hub Height (m)	92	90
Rotor Diameter (m)	126	116

Table 4
Sensor types and their descriptions with physical units.

Sensor type	Description	Physical unit
DMS	Strain gauge	kNm
FOS	Fiber Optic Sensor	λ (wavelength)
Environmental	Temperature, humidity, pressure	°C, %, hPa
Inclinometer	Inclination angle	°
Control	Various	rpm, °, kW, m/s
ICP	Accelerometer	m/s^2

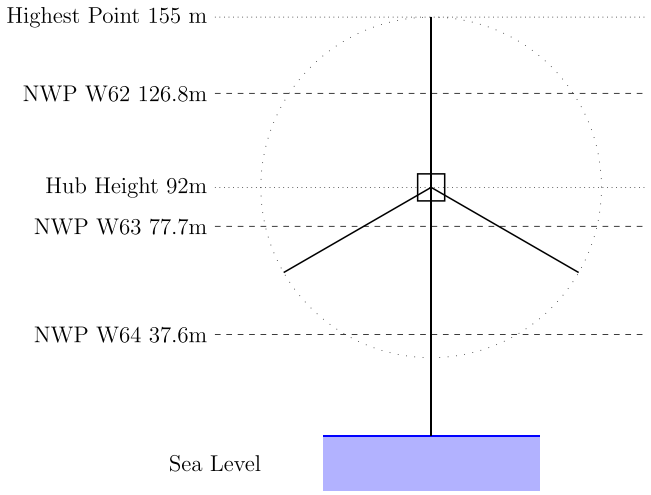


Fig. 2. Schematic of an offshore wind turbine with NWP model height levels. W63 variable is the nearest to the center of the turbine.

4. Experimental setup

This section presents the experimental setup used for developing the wind power forecasting models. It begins with a brief overview of the selected offshore wind farm, and presents RAVE and Pamore, the databases used to train the models. To address missing values, a data recovery strategy was applied using meteorological data, XGBoost, and the physical wind power equation. Outliers were identified and filtered using the Isolation Forest method. The final dataset was constructed by selecting relevant variables, followed by normalization and splitting into training, validation, and test sets.

4.1. Case study

The Alpha Ventus offshore wind farm, situated in the North Sea off the coast of Germany, 60 km from the coast, has been operational since 2010. This farm consists of 6 Adwen AD-5116 turbines of 5 MW with tripod foundations, and 6 Senvion 5M turbines with jacket foundations, totaling 60 MW of installed capacity. Table 3 shows the most important characteristics of the two types of turbines. The water depth in the area is about 30 m. The wind farm is managed by the enterprises EWE, RWE, and Vattenfall. FINO1 is a research platform built at *N* 54° 00′ 53,5” E 6° 35′ 15,5” in the vicinity of the farm. The study period spans from September 1, 2022, to July 9, 2023, a duration of 10 months.

4.2. RAVE data

Data from FINO1 and wind turbine sensors can be accessed through *Research at Alpha VEntus* (RAVE) platform. Although the sensors have different measurement frequencies, ten-minute aggregated data are available to download from the RAVE platform. Sensor data from the Senvion 5M turbine AV04 was used, as it has the longest measurement history and is the closest turbine to the FINO1 platform, making the platform’s meteorological data more representative of the turbine’s operating conditions. Turbine sensors are located in the foundations, tower, nacelle, blades, and rotor shaft. Table 4 shows the different turbine sensor types.

We used data from the nacelle to train the models. Only wind speed (m/s), generator speed (rpm), pitch angle (°), Azimuth angle (°), temperature (° C), and atmospheric pressure (hPa) have a correlation over 0.2 in absolute value and were kept. We observe that the Azimuth angle, pitch angle, and generator speed are strongly correlated. We also included measurements from FINO1 of atmospheric pressure (hPa) at 20 m and wind speed (m/s) at 40, 50, 70, 80, 90, and 100 m.

4.3. Numerical weather prediction

Numerical Weather Prediction (NWP) models are mathematical models based on physical principles to predict the weather. The German Weather Service (Deutscher Wetterdienst, DWD) operates a regional model, ICON-D2, which is limited to Germany and its neighboring countries, and is primarily used for very short-range forecasting up to 48 h at a horizontal resolution of 2.2 km [14,15]. The model forecasts atmospheric conditions at 65 atmosphere levels. The data is available online through Pamore (PArallel MOdel data REtrieve from Oracle databases) platform. Since wind speed is part of the theoretical wind power equation, we have used predicted data for wind speed at layers 62, 63, and 64, corresponding to heights of 126.8 m, 77.7 m, and 37.6 m above sea level, respectively [41]. These heights are those where the wind interacts with the wind rotor and are represented in Fig. 2.

4.4. Data recovery

For atmospheric variables at the nacelle, we employed a two-step process to address missing data. Initially, missing values in the FINO1 research platform dataset were filled using data from the meteostat Python package [42]. The filled FINO1 data was then used as input to XGBoost models to predict and recover missing data from nacelle sensors.

For atmospheric pressure at the nacelle, missing values were imputed using pressure data measured at 20 m from FINO1. For wind speed at the nacelle, data from multiple heights (100, 90, 80, 70, 60, 50, and 40 m) were used as inputs to predict and recover missing values

Table 5
Key variables, their units, sources, and filling methods. Filling methods applied to RAVE data apply physical formulae and XGBoost models trained with other variables.

	Name	Units	Source	Filling method
AP	Active Power	kW	RAVE	Theoretical wind power equation
WS	Wind speed at the nacelle	m/s	RAVE	XGBoost using FINO1 data
APr	Atmospheric pressure at the nacelle	hPa	RAVE	XGBoost using FINO1 data
GS	Generator speed	rpm	RAVE	XGBoost using azimuth angle and rotor position
W62	Wind speed at 126.8 m	m/s	NWP	–
W63	Wind speed at 77.7 m	m/s	NWP	–
W64	Wind speed at 37.6 m	m/s	NWP	–

at the nacelle. Similarly, temperature data at 100 m were employed to recover missing values for the temperature at the nacelle. Azimuth angle and rotor position were used as predictors in an XGBoost model to recover missing generator speed values.

Finally, missing values for the objective variable, active power, were addressed using the theoretical wind power equation:

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

where P_w is the wind power (W), A is the area covered by the rotor blades (m^2), ρ is the air density (kg/m^3), v is the wind speed at the nacelle (m/s), and C_p is the turbine power coefficient. The area was calculated using the formula for the area of a circle with the rotor radius. Air density was calculated as:

$$\rho = PRT \tag{3}$$

where P is the atmospheric pressure (Pa), R is the ideal gas constant (287.05 J/(kg K)), and T is the temperature (K). The turbine power coefficient C_p, which is not publicly available, was estimated as the value that minimized the mean squared error (MSE) between calculated and observed wind power in non-missing entries. Table 5 shows a summary of the curated variables used in the models.

4.5. Outlier detection and treatment

The ideal power curve of a wind turbine represents the expected relationship between wind speed and power output under normal operating conditions. This curve is characterized by three main regions: a low wind speed region where no power is generated, an intermediate region of increasing power output, and a plateau where the turbine operates at its rated capacity. However, deviations from this curve, often referred to as outliers, can occur due to various factors such as turbine curtailment, measurement errors, or operational anomalies.

As shown in Fig. 3, outliers were classified into three distinct types based on their characteristics. Type I outliers were slightly negative values resulting from the power consumption required to initiate rotor movement. These values were transformed to zero, as they do not contribute to wind power output. Type II outliers corresponded to intermediate power values observed during turbine shutdown phases. The horizontal lines (Type III) in the power-wind curve result from curtailment operations, which involve the intentional reduction of electricity output from wind turbines. Grid operators typically carry out these measures to ensure grid stability or to perform maintenance.

To detect Type II and Type III outliers, we employed the Isolation Forest algorithm, a machine learning-based anomaly detection method known for its effectiveness in identifying deviations in high-dimensional datasets. The detection results are illustrated in Fig. 4(a), where the identified anomalies are marked in red. Once detected, Type II and Type III outliers were replaced using the theoretical wind power equation. Fig. 4(b) shows the cleaned dataset after imputation.

4.6. Dataset configuration

The dataset is composed of data from the wind turbine measurements and the NWP simulations. All the data is hourly sampled. The

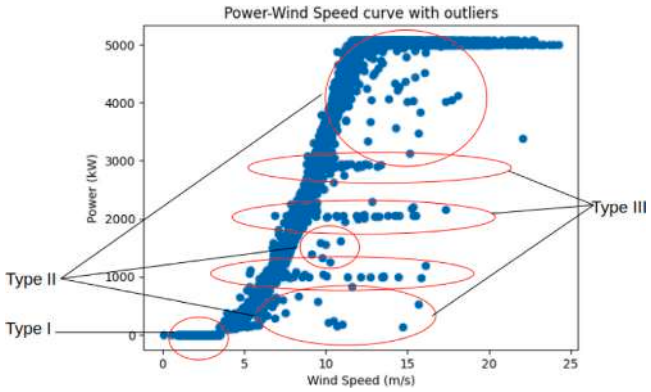


Fig. 3. Outliers types in the observed power curve of the turbine. Type I are caused during turbine start, type II by intermediate shut down phases and Type III by curtailment operations.

measured variables include active power, wind speed, atmospheric pressure and generator speed. The wind speed has been simulated at instant t+1 with the information at instant t, at three different heights corresponding to the layers 62, 63, and 64. The variable to predict is the following hour's value of the active power. Note that historical values of the active power can be used as an explanatory variable.

The data notation is as follows:

- y_{t+1} is the objective variable, the active power at time $t+1$.
- $x_{WS,t}$, $x_{AP,t}$, $x_{APr,t}$, and $x_{GS,t}$ are the explanatory variables measured at instant t: wind speed, active power, atmospheric pressure, and generator speed. Note that $x_{AP,t}$ is the same as y_t. $x_{XX,t}$ will be used to represent any of these variables.
- $\hat{x}_{NWP62\ t+1|t}$, $\hat{x}_{NWP63\ t+1|t}$ and $\hat{x}_{NWP64\ t+1|t}$ are the NWP simulation at instant t for the time horizon $t+1$ of the wind speed in the layer 62, 63 or 64. $\hat{x}_{NWPXX\ t+1|t}$ will be used to represent any of these variables.

4.7. Dataset preparation

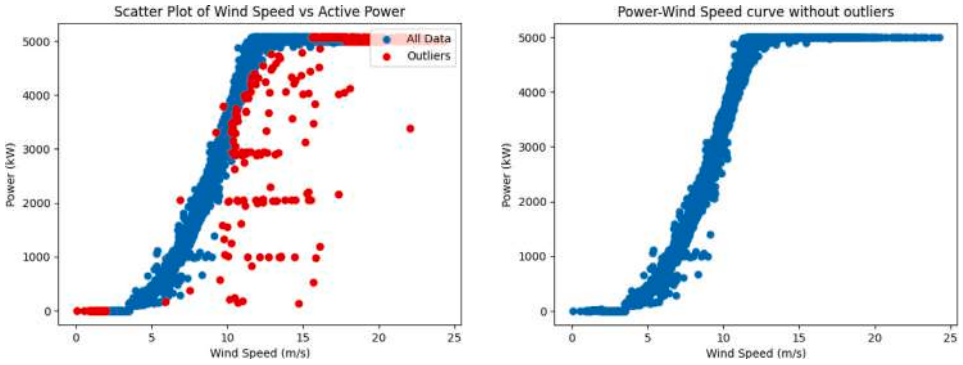
To ensure consistency and comparability, all input variables were normalized prior to model training. Normalization was performed using the formula:

$$x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where:

- x is the original value of the variable.
- $x_{\min}$ is the minimum value of the variable in the dataset.
- $x_{\max}$ is the maximum value of the variable in the dataset.
- $x_{\text{normalized}}$ is the scaled value, ensuring the variable falls within the range [0, 1].

The objective variable for all experiments was the wind power generated in the next nine hours, resulting in nine output values. 80%



(a) Observed power curve with outliers marked in red.

(b) Cleaned power curve after outlier removal.

Fig. 4. Comparison of the observed and cleaned power curves. (a) shows the raw observed data with outliers highlighted, while (b) displays the cleaned data after preprocessing and imputation.

of the data was allocated to training, an additional 10% of the data was designated for the adjustment of hyperparameters, and the remaining 10% was reserved as the test set to assess the overall effectiveness and generalization capability of the trained models.

4.8. Experiments configuration

This paper aims to analyze the impact of different elements on short-term forecasting. Therefore, all the experiments predict the active power for the following hour. The contribution of different combinations of measured variables to the performance improvement were studied both with and without the presence of the different NWP variables, using an LSTM model. Then, the impact of including lags of the explanatory variables were compared with the cases without any lag. Finally, the experiments are repeated with an MLP model to measure the utility of using the LSTM, a more computationally expensive model. Both LSMT and MLP model hyperparameters are adjusted using the Optuna framework, as explained in Section 3.3. The experiments are as following:

1. Compare the performance of LSTM models using different combinations of explanatory variables $f(x_{\mathrm{XX},t}, \ldots, x_{\mathrm{XX},t})$ with models that include NWP values $f(x_{\mathrm{XX},t}, \ldots, x_{\mathrm{XX},t}, \hat{x}_{\mathrm{NWPXX\ t+1|}t})$ for the three layers: 62, 63, 64.
2. Analyze the contribution of n lags of the feature variables in LSTM models, such as $f(x_{\mathrm{XX1},t}, x_{\mathrm{XX2},t}, \ldots, x_{\mathrm{XX1},t-n}, x_{\mathrm{XX2},t-n})$ compared to models using only current inputs, i.e., $f(x_{\mathrm{XX1},t}, x_{\mathrm{XX2},t})$. Additionally, study both cases with the inclusion of NWP variables.
3. Repeat the experiments using MLP, XGBoost, Random Forest, and LightGBM algorithms to determine whether the same behavior is observed.
4. Compare the performance of baseline and LSTM models to justify the use of more computationally expensive algorithms such as LSTM.

5. Results and discussion

This section presents the evaluation of the predictive performance of different model configurations for short-term wind power forecasting. The analysis focuses on the contribution of numerical weather prediction (NWP) data when combined with the explanatory variables wind speed, active power, generator speed, and atmospheric pressure. Some combinations were also tested with different lag lengths. The Mean Absolute Error (MAE) was used as the evaluation metric. Additionally, four different machine learning methods were used as a reference to benchmark the performance of the LSTM-based models: multilayer perceptron (MLP), Random Forest, XGBoost and LightGBM. The statistical significance of the observed differences in forecasting accuracy was examined using the Diebold–Mariano test [43].

5.1. Impact of explanatory and NWP variables

Here, we evaluate the impact of combining different explanatory variables with numerical weather prediction (NWP) data. Table 6 reports the mean absolute error (MAE) obtained for each model configuration. The results show that, with the exception of the configuration using all four explanatory variables, adding an NWP variable can improve prediction accuracy in at least one model for each group of explanatory variables. The best result was achieved by combining wind speed, active power, and air pressure with the NWP variable W62, as shown in Fig. 5.

To assess the statistical significance of the observed improvements, Table 6 also marks with an asterisk (*) the cases where the improvement is significant after the Diebold–Mariano test, compared with the corresponding model without any NWP variable. A significance threshold of 0.05 was used.

Among the NWP variables tested, W63 showed a statistically significant impact in 5 out of the 7 configurations, W62 in 4 configurations, and W64 in three. These results highlight the relevance of selecting the appropriate NWP input depending on the available explanatory variables and the forecasting horizon.

It is worth noting that the ten best-performing models all included active power as one of the explanatory variables. This is a logical outcome, as active power is the same variable that the models aim to predict, and its recent values are likely to carry relevant information for short-term forecasting.

5.2. Impact of lagged variables

This section explores the effect of incorporating lagged values of input variables on the performance of wind power prediction models. As we are dealing with time series data, incorporating temporal dependencies through lagged inputs may provide valuable context to improve forecasting accuracy.

To assess this, we tested the inclusion of 1, 2, 3, 4, 6, and 12 lags for both explanatory and NWP variables. The experiments were conducted using the combinations of explanatory variables that showed the best performance in the previous analysis.

Table 7 presents the MAE obtained for each configuration across varying numbers of lags, different sets of explanatory variables, and with or without NWP inputs. In the table, the results where adding lags

Table 6
MAE values (kW) for the LSTM model with and without NWP. The values marked with * indicate a significant difference (p-value<0.05) between the models. The p-values are calculated using the Diebold–Mariano test. The best result is in bold. While W63 shows the most consistent improvements, the best model combines W62 with Wind Speed, Active Power and Air Pressure.

Explanatory variables	No NWP	W62	W63	W64
AP, APr	234.63	229.49 *	236.27	245.94
AP	235.79	250.05	232.68 *	233.72
WS, AP, APr	243.11	**228.93** *	231.10 *	239.25 *
WS, AP	246.73	234.67 *	239.11 *	241.89 *
WS, APr, GS, AP	236.91	239.21	239.19	242.40
WS, APr, GS	260.13	263.09	243.21 *	260.88
WS	265.54	245.78 *	246.71 *	259.83 *

significantly improved the model are marked with a dagger (†). The model was compared to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. These improvements were validated using the Diebold–Mariano test with a p-value threshold of 0.05.

The results indicate that adding a single lag tends to improve performance in most cases. However, further increasing the number of lags does not result in additional benefits. The best overall performances are achieved with models that include one lag and explanatory variables such as wind speed, active power, and air pressure. Generator speed appears to contribute less consistently to prediction accuracy.

Table 7 also assesses the statistical significance of the improvement provided by adding NWP variables to models with the same explanatory variables and number of lags. The cases where a significant improvement was achieved were marked with an asterisk (*). Among the NWP variables tested, W63 demonstrated the most consistent and significant impact across experiments.

Focusing on the experiments with one lag, W63 significantly improved the results in all cases, while W62 and W64 each showed statistically significant contributions in four and three out of five configurations, respectively. These findings confirm the benefit of incorporating temporal information, particularly when used in conjunction with relevant NWP variables.

5.3. Results across conventional ML models

This section examines the impact of incorporating NWP predictors and lagged explanatory variables when employing conventional machine learning models. The experiments are conducted with Multi-Layer Perceptron (MLP), XGBoost, Random Forest, and LightGBM, and are replicated both with and without NWP variables across different lag lengths. Complete results for all configurations are provided in Appendix.

Table 8 reports the percentage of cases in which adding one or more lags yields a statistically significant improvement according to the Diebold–Mariano test, relative to the relevant model with fewer lags. Overall, adding a single lag leads to consistent gains across all model families, with the highest incidence of significant improvement observed for MLP (75% of cases) and the lowest for LightGBM (46% of cases).

By contrast, including additional lags beyond the first appears to provide limited incremental benefit for most technologies. MLP is the only model showing a noticeable sensitivity to longer lag structures, where the inclusion of specific additional lags can lead to further improvements in up to 21% of cases.

Table 9 summarizes the percentage of cases in which the inclusion of the three NWP variables yields a statistically significant improvement relative to the identical configuration without NWP predictors. For consistency and interpretability, this analysis is restricted to the 0-lag and 1-lag specifications. The inclusion of NWP variables improves performance consistently across all technologies. The lowest incidence of significant improvement occurs for MLP (50% of cases). Finally, no single NWP predictor consistently dominates the others in terms of the frequency of significant improvements (see Fig. 5).

5.4. Comparison between forecasting models

Long Short-Term Memory (LSTM) networks are generally more complex than the alternative model classes considered in this study and, as a result, require higher computational resources and longer training and inference times.

The overall ranking of the best-performing configurations further highlights the competitiveness of LSTMs in this application. Among the 50 best models, 36 correspond to LSTM variants, followed by 12 MLP configurations and 2 LightGBM configurations; moreover, the top 10 models are all LSTM-based.

Table 10 presents the top 5 best performing configurations per model family. In terms of the best observed performance, the leading configuration is an LSTM model with an error of 224.22 kW. The best non-recurrent alternatives achieve slightly higher errors, with the best MLP at 232.56 kW, LightGBM at 236.25 kW, XGBoost at 242.11 kW, and Random Forest at 243.63 kW. All best performing models use NWP variables and most include 1-h lagged variables.

Overall, LSTM models consistently deliver the strongest predictive performance in this benchmark. Nevertheless, the remaining algorithms also provide competitive results and may be preferable in use cases where computational cost, implementation complexity, and operational constraints outweigh marginal gains in accuracy.

6. Feature attribution using SHAP values

This section complements the accuracy metrics by providing a model-based attribution analysis aimed at understanding how Numerical Weather Prediction (NWP) features influence short-term power forecasts. Shapley Additive Explanations (SHAP) provide a principled feature attribution where a prediction is decomposed into additive contributions from each input feature relative to a baseline expectation. For a given instance, each SHAP value quantifies how much a feature increases or decreases the model output [44].

To analyze the impact of NWP predictors, we employ XGBoost models because of its good performance and its direct integration with SHAP. We select the top 10 performing XGBoost models using each of the three NWP variables and compute SHAP values for the corresponding feature sets.

Table 11 summarizes the contribution of the NWP variables, including both the contemporaneous value and the one-hour lag (denoted by the suffix H-1). Across the analyzed models, the mean absolute SHAP value associated with the NWP variables ranges from 95 to 130, which is significant when forecasting a target with a median value of 674. In relative terms, the mean SHAP contribution of NWP features corresponds to approximately 4%–6% of the total attribution magnitude across all explanatory variables, indicating that NWP predictors are not negligible drivers of the forecast. Lagged NWP predictors are, on average, slightly less influential than their contemporaneous counterparts but still contribute meaningfully.

The variability of the lagged-attribution distributions is substantial: the standard deviation of the lagged NWP SHAP values is typically between 25% and 50% of the standard deviation observed for the corresponding non-lagged NWP feature, consistent with a secondary yet complementary informational role.

The dependence patterns in Fig. 6 provide additional insight into how NWP variables affect the forecast. The contemporaneous NWP features tend to have limited impact at low values, reduce the predicted output for median values, and increase the predicted output for large values. This behavior is consistent with a corrective role in the tails, where forecasting is typically more difficult, and errors are larger.

Table 7
MAE values (kW) for the **LSTM** model with and without NWP. Values marked with † denote a statistically significant improvement (p-value <0.05) relative to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. Values marked with * indicate a significant improvement (p-value <0.05) compared to the No NWP counterpart. The p-values are calculated using the Diebold–Mariano test. Although using one-hour lagged variables yields to significant improvement, more lags does not lead to further improvement. The inclusion of any of the NWP variables leads to significant improvements.

Explanatory variables	NWP	0	1	2	3	4	6	12
AP, APr	No NWP	234.63	239.46	251.42	249.16	246.40	254.27	319.58
AP, APr	W62	229.49	231.13	241.47 *	247.58 *	895.69	247.45 *	902.51
AP, APr	W63	236.27	226.82 *	236.72	258.13	242.07	251.84 *	243.01
AP, APr	W64	245.94	231.10 †	240.92	245.60 *	249.56	258.56	901.46
AP	No NWP	235.79	237.44	239.21	239.83	245.71	251.42	249.94
AP	W62	250.05	228.81 * †	243.64	250.03	249.05	244.19	318.84
AP	W63	232.68 *	230.20 * †	242.23	245.85	243.60 *	251.07 *	902.91
AP	W64	233.72 *	234.77 *	241.35	252.38	252.53	288.89	267.74
WS, AP, APr	No NWP	243.11	233.28 †	248.52	250.13	241.59	253.03	902.78
WS, AP, APr	W62	228.93 *	226.75 * †	241.10 *	252.02	243.29	251.06 *	253.05 *
WS, AP, APr	W63	231.10 *	**224.22** * †	238.37 *	252.89	251.78	901.70	902.27 *
WS, AP, APr	W64	239.25 *	234.92 †	240.06 *	248.29 *	246.38	897.17	257.50 *
WS, AP	No NWP	246.73	235.13 †	239.31	246.86	242.83	262.25	249.90
WS, AP	W62	234.67 *	225.38 * †	234.04 *	241.26 *	248.85	247.05	261.79
WS, AP	W63	239.11 *	230.26 * †	244.67	247.76	254.53	244.16 *	246.86 *
WS, AP	W64	241.89 *	230.63 * †	245.72	250.40	243.31	251.94 *	254.66
WS	No NWP	265.54	267.33	275.15	281.74	276.92	897.99	902.61
WS	W62	245.78 *	256.14 *	262.23 *	263.72 *	274.60 *	901.39	272.35 *
WS	W63	246.71 *	256.02 *	257.01 *	267.17 *	271.84 *	257.96 *	344.45
WS	W64	259.83 *	255.95 *	254.78 *	259.83 *	261.49 *	273.24 *	285.14 *

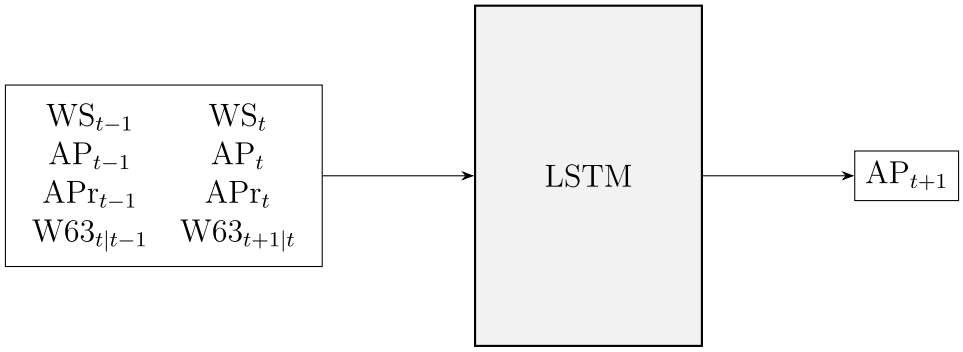


Fig. 5. Schema of inputs and outputs of the best forecasting model. The explanatory variables include Wind Speed, Active Power and Air Pressure combined with NWP variable W63. All variables are included both current and 1-h lagged. The algorithm used is LSTM to forecast Active Power.

Table 8
Proportion of Diebold–Mariano tests that resulted in statistically significant improvements (p-value <0.05) when increasing the number of lags used as input features for each model. The table shows how including 1-h lagged variables yields improvements in the majority of cases, while more lagged variables do not seem relevant, except for the MLP algorithm.

	LSTM	MLP	Random forest	XGBoost	LightGBM
1 lags	0.55	0.75	0.61	0.57	0.46
2 lags	0.00	0.18	0.00	0.07	0.14
3 lags	0.00	0.21	0.07	0.00	0.04
4 lags	0.00	0.18	0.04	0.00	0.07
6 lags	0.00	0.21	0.00	0.07	0.00
12 lags	0.00	0.11	0.00	0.00	0.00

Table 9
Proportion of Diebold–Mariano tests that resulted in statistically significant improvements (p-value <0.05) when including NWP wind speed as an input feature for each model. The inclusion of NWP variables yield to significant improvements in most of cases.

NWP	Lags	LSTM	MLP	Random forest	XGBoost	LightGBM
W62	0 lag	0.60	0.43	0.71	0.86	0.86
	1 lag	0.80	0.57	1.00	1.00	0.86
W63	0 lag	0.80	0.43	1.00	0.86	0.86
	1 lag	1.00	0.57	0.71	1.00	0.86
W64	0 lag	0.80	0.29	0.86	0.86	0.71
	1 lag	0.60	0.71	1.00	1.00	1.00

By contrast, the one-hour lagged NWP features exhibit an opposite distributional pattern, suggesting that lagged information can partially counterbalance or refine the contemporaneous NWP contribution and help calibrate the net impact of NWP inputs across the feature range.

7. Performance time and computational resources

This section characterizes the computational resources and execution times associated with the proposed forecasting pipeline, with the objective of supporting reproducibility and informing deployment-oriented trade-offs between accuracy and operational cost.

All experiments were executed on a CPU-only Linux server (kernel 5.15.0-101-generic, x86_64, glibc 2.31) equipped with an Intel(R) Core(TM) i9-14900KF processor (single socket) and 125 GiB of RAM. Although the system provides 32 logical CPUs, each experiment was run using a single processor to ensure consistent and comparable runtime measurements across model configurations.

For every model, hyperparameters were tuned using Optuna with 150 trials per configuration. Table 12 reports the percentile statistics of the duration required to complete a 150-trial for a specific technology and variables configuration. Three runtime regimes emerge. First,

Table 10
MAE results (kW) for the top 5 performing configurations by model type. LSTM stands out with the best results, followed by the MLP. All models use NWP variables and most include 1-h lagged explanatory variables.

	LSTM				MLP			
	MAE	Vars	NWP	Lags	MAE	Vars	NWP	Lags
1	224.22	WS, AP, APr	W63	1	232.56	WS, APr, GS, AP	W63	1
2	225.38	WS, AP	W62	1	235.32	AP	W62	1
3	226.75	WS, AP, APr	W62	1	236.48	AP, APr	W64	3
4	226.82	AP, APr	W63	1	238.45	AP, APr	W63	4
5	228.81	AP	W62	1	238.45	WS, APr, GS, AP	W62	2
	Random Forest				XGBoost			
	MAE	Vars	NWP	Lags	MAE	Vars	NWP	Lags
1	243.63	AP, APr	W62	0	242.11	AP, APr	W63	0
2	243.82	WS, AP, APr	W63	0	242.97	WS, APr, GS, AP	W62	1
3	245.94	AP, APr	W63	0	244.11	AP, APr	W62	1
4	246.35	WS, AP, APr	W63	1	244.51	AP	W62	2
5	246.95	AP, APr	W62	1	244.55	AP	W62	1
	LightGBM							
	MAE	Vars	NWP	Lags				
1	236.25	AP, APr	W63	1				
2	236.32	AP, APr	W62	1				
3	240.98	WS, APr, GS, AP	W63	1				
4	241.10	WS, APr, GS, AP	W63	0				
5	241.97	WS, AP, APr	W64	0				

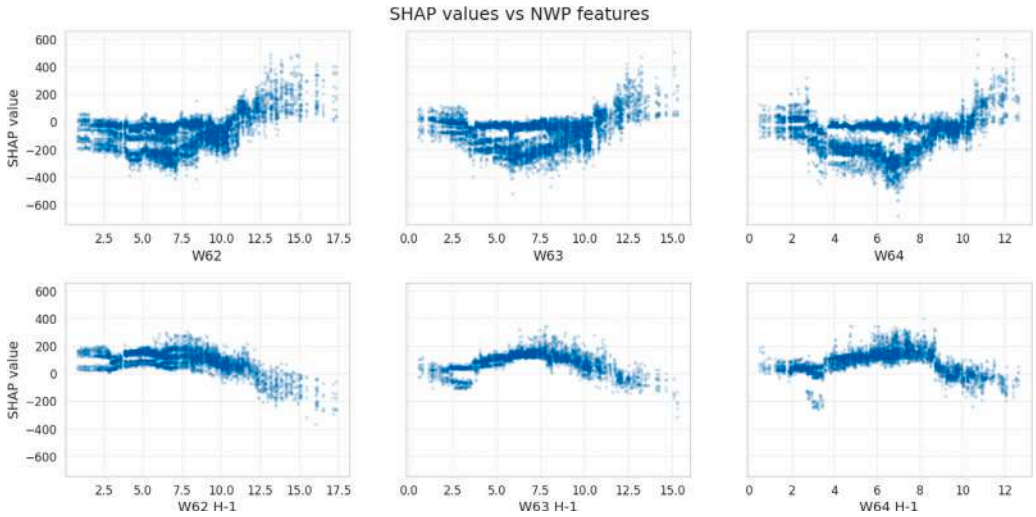


Fig. 6. Scatter plots NWP wind speed features against its corresponding SHAP values for the top-10 models per wind speed NWP source. Lagged variables are marked with H-1 and show a behavior opposite to the not-lagged variables.

Table 11
Summary of Wspeed NWP features' importance across the top-10 models per windspeed NWP source. Importance is measured as the mean absolute SHAP value, and its share relative to the total mean absolute SHAP values across all features. Not lagged variables show slightly greater importance than lagged variables, and more variability.

NWP	Number of models	Mean |SHAP|	Std |SHAP|	Mean relative SHAP
W62	10	117.9	69.1	0.057
W62 H-1	6	106.3	38.5	0.05
W63	10	103.5	65.4	0.048
W63 H-1	5	99.5	15.4	0.044
W64	10	127.3	78.9	0.059
W64 H-1	6	102.5	20.4	0.046

gradient-boosted tree models (XGBoost and LightGBM) are the fastest, with 90% of searches completing between 20 s and 2 min. Second, MLP and Random Forest exhibit intermediate-to-high cost, with 90% of searches spanning approximately 4 min to 1 h 11 min. Third, LSTM hyperparameter searches are substantially more expensive, with 90% of searches requiring between 3 h and 11 h.

Once hyperparameters are selected, the operationally relevant cost reduces to model retraining time and inference latency. Table 13 reports percentile statistics for LSTM training and inference. In this setting, LSTM training time lies between 1 and 9 min for 90% of cases, while the maximum inference time is 2 s for 7248 entries, indicating that inference has a limited impact on end-to-end latency.

Given the markedly higher cost of LSTM hyperparameter optimization, a practical strategy is to identify a competitive hyperparameter set and retrain the model regularly while keeping hyperparameters fixed, performing full searches less frequently. Conversely, for faster model families (in particular XGBoost and LightGBM), hyperparameter searches can be repeated more regularly, depending on the operational use case and the available computational resources. Overall, inference time is negligible compared with training and hyperparameter optimization.

8. Conclusions and future work

This study conducted a series of experiments on an A5 MW turbine from the offshore wind farm Alpha Ventus. We analyzed the impact of different explanatory variables, the inclusion of numerical weather prediction (NWP) variables, and the use of lagged variables to predict next-hour wind power output. The performance was evaluated using a state-of-the-art recurrent neural network (Long Short-Term

Table 12
Training bayesian optimization duration percentiles by technology. LSTM stands out as the most computationally expensive algorithm to train. Gradient Boosting algorithms (XGBoost and LightGBM) are the fastest training algorithms.

	Total of runs	5th pct.	25th pct.	50th pct.	75th pct.	95th pct.
LSTM	148	3 h 15 m	5 h 5 m	7 h 9 m	8 h 53 m	11 h 6 m
LightGBM	196	32 s	38 s	46 s	54 s	1 m 30 s
MLP	196	5 m 2 s	8 m 40 s	12 m 11 s	16 m 21 s	29 m 7 s
Random Forest	196	4 m 56 s	10 m 40 s	20 m 23 s	35 m 56 s	1 h 11 m
XGBoost	196	24 s	29 s	44 s	1 m 1 s	1 m 56 s

Table 13
LSTM training and inference time percentiles. The inference time is calculated for a batch of 7248 entries.

	Training time	Inference time
5th pct.	1 m 1 s	1 s
25th pct.	2 m 0 s	1 s
50th pct.	3 m 8 s	2 s
75th pct.	4 m 29 s	2 s
95th pct.	8 m 9 s	2 s

Memory, LSTM) alongside four established machine learning baselines: Multi-Layer Perceptron (MLP), XGBoost, Random Forest, and LightGBM.

A data recovery process was implemented using the XGBoost algorithm and the theoretical wind power formula. Additionally, outliers in the dataset were identified and filtered using the Isolation Forest method. The NWP model employed in this study was ICON-2 (ICOsahedral Nonhydrostatic D2), a publicly available and high-resolution physical model. The outputs used from this NWP model are the wind speed at three different heights (W62, W63 and W64).

Regarding explanatory variables, different combinations yielded better results depending on the specific experiment. Nevertheless, active power, the same variable targeted for prediction, proved to be useful in nearly all top-performing models. Wind speed and air pressure also played important roles in improving prediction performance.

The inclusion of NWP variables significantly enhanced model accuracy, with the best results obtained when NWP variables were included. Among them, W63 consistently led to the greatest improvements for the LSTM algorithm. This variable corresponds to the vertical level closest to the center of the wind turbine, which may explain its relevance. For the rest of algorithms, the inclusion of any of the three NWP variables led to significant improvement. In terms of lagged variables, although using one-hour lagged variables often yielded to significant performance improvement, including more lags seldom led to further significant improvements.

Finally, the LSTM architecture outperformed the baseline models. Despite its greater training complexity, the LSTM consistently delivered significantly better predictions, justifying the additional computational effort depending on the use case and the available resources.

These results demonstrate the effectiveness of the proposed forecasting framework for the Alpha Ventus offshore wind farm. This enabled a controlled and detailed assessment of the interactions between explanatory variables, NWP inputs, lagged information, and different machine learning architectures. However, extending these findings to a broader international context is a crucial next step to ensure their general applicability. As offshore wind deployments continue to expand globally and higher-resolution NWP models and measurement systems generate increasingly rich datasets, machine learning is expected to play an even more prominent role in wind power forecasting. Future research should therefore evaluate the transferability of the proposed methodology across offshore wind farms located in different geographical and climatic regions by incorporating globally recognized NWP systems such as the Global Forecast System (GFS) and the European Centre for Medium-Range Weather Forecasts (ECMWF) model. Applying the framework to sites with diverse turbine technologies, hub heights, and meteorological regimes, such as those found along the coasts of North America or East Asia would enable a rigorous assessment of model robustness, NWP feature relevance, and scalability. Moreover, the integration of advanced deep learning architectures, including attention-based, physics-informed, probabilistic, and transferable forecasting approaches, represents a promising direction for developing more universally robust wind power forecasting solutions capable of supporting the reliable large-scale integration of offshore wind energy into future power systems.

CRediT authorship contribution statement

Eloy Insunza: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Carlos de los Santos Jiménez:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Antonio Muñoz:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **José Portela:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Ibuki Kusano:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Horacio Rostro-Gonzalez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve text cohesion and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Antonio Munoz San Roque reports financial support was provided by Ramon Llull University and Comillas Pontifical University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge RAVE (Research at Alpha Ventus) for providing access to wind turbine and weather data, which was essential to the successful completion of this research.

Appendix. MLP, Random Forest, XGBoost, and LightGBM results

See Tables A.14–A.17.

Table A.14
MAE values (kW) for the **MLP** model with and without NWP. Values marked with † denote a statistically significant improvement (p-value<0.05) relative to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. Values marked with * indicate a significant improvement (p-value<0.05) compared to the No NWP counterpart. The p-values are calculated using the Diebold–Mariano test. Using one-hour lagged variables yields to significant improvement, but adding more lags lead to further improvement in a few cases. The inclusion of any of the NWP variables leads to significant improvements.

Explanatory variables	NWP	0	1	2	3	4	6	12
AP, APr	No NWP	255.96	241.64 †	249.59	245.89	250.58	247.52	249.36
AP, APr	W62	250.65 *	244.25	243.83	239.95	246.37	243.14 * †	245.34 *
AP, APr	W63	239.43	238.48 †	243.39 *	240.85 *	238.45 * †	242.82 *	241.85 *
AP, APr	W64	1075.33	241.14 * †	247.87	236.48	238.73 *	243.25 *	241.55 *
AP	No NWP	265.56	249.22 †	251.07	253.34	247.72 †	260.83	254.52
AP	W62	243.45 *	235.32 * †	240.01 *	253.41	246.41 *	242.31 *	244.01 *
AP	W63	244.03 *	242.21 * †	243.76 *	243.05 *	245.59 *	242.27 *	248.04 *
AP	W64	245.10 *	240.73 * †	262.14	250.76 *	257.04	243.24 *	240.69 * †
WS, AP, APr	No NWP	253.87	240.96 †	248.26	256.06	247.53	251.17	248.57
WS, AP, APr	W62	248.99 *	242.31 †	243.10 *	246.37 *	242.14 * †	239.50 * †	247.15 *
WS, AP, APr	W63	251.72 *	245.77 †	248.64	258.71	245.12 * †	240.93 * †	244.84 *
WS, AP, APr	W64	251.02 *	242.34	243.63 * †	241.11 * †	252.83	242.74 *	246.69 *
WS, AP	No NWP	250.52	257.01	252.77	250.47 †	250.97	246.39 †	247.90
WS, AP	W62	261.71	244.04 †	247.98 *	253.15	241.61 * †	247.24	246.69 *
WS, AP	W63	253.51	239.16 * †	239.36 *	242.92 *	243.61 *	244.66 *	253.90
WS, AP	W64	258.72	245.89 * †	244.69 * †	246.34 *	249.03 *	243.04 * †	241.73 * †
WS, APr, GS, AP	No NWP	246.52	250.90	253.29	246.52 †	252.71	261.06	258.69
WS, APr, GS, AP	W62	252.26	242.77 * †	238.45 * †	260.49	254.40	240.80 *	241.21 *
WS, APr, GS, AP	W63	262.16	**232.56** * †	241.95 *	244.82 *	246.93 *	243.88	247.19 *
WS, APr, GS, AP	W64	247.57	250.47 *	241.84 * †	245.08 *	249.73	250.57 *	244.45 *
WS, APr, GS	No NWP	290.03	292.15	281.59 †	324.24	293.88	286.63	290.41
WS, APr, GS	W62	293.87	276.06 * †	276.34 *	285.19	261.83	298.77	288.70 *
WS, APr, GS	W63	287.43	270.86 †	296.98	262.35 * †	290.77 *	288.85	261.95 * †
WS, APr, GS	W64	438.83	387.82 †	432.44	272.89 * †	277.93 *	270.38 * †	274.62 *
WS	No NWP	277.23	304.74	284.28	277.01 †	327.75	304.16	282.49
WS	W62	415.02	273.01 * †	278.72 *	284.74	418.26	290.00 *	290.95
WS	W63	272.74 *	263.51 * †	294.59	265.08 *	281.62 *	418.41	381.68
WS	W64	296.83	254.74 * †	288.87	283.73	410.23	281.03 *	288.91

Table A.15
MAE values (kW) for the **Random Forest** model with and without NWP. Values marked with † denote a statistically significant improvement (p-value<0.05) relative to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. Values marked with * indicate a significant improvement (p-value<0.05) compared to the No NWP counterpart. The p-values are calculated using the Diebold–Mariano test. Although using one-hour lagged variables yields to significant improvement, more lags does not lead to further improvement. The inclusion of any of the NWP variables leads to significant improvements.

Explanatory variables	NWP	0	1	2	3	4	6	12
AP, APr	No NWP	261.15	259.24 †	261.12	258.88 †	259.32	260.61	260.12
AP, APr	W62	243.63	246.95 *	249.23	249.40 *	249.65 *	252.54 *	253.67 *
AP, APr	W63	245.94 *	249.24	250.78	251.14	251.74 *	252.74 *	255.89 *
AP, APr	W64	247.34	252.18 *	254.14 *	254.28 *	252.99 *	255.11 *	256.28 *
AP	No NWP	271.23	259.59 †	260.52	258.49 †	259.17	264.68	263.68
AP	W62	249.01 *	248.48 * †	248.67 *	251.43 *	250.03 *	255.64 *	257.35 *
AP	W63	253.68 *	250.73 * †	253.91 *	253.91	254.31 *	258.30 *	258.37 *
AP	W64	260.18 *	256.83 * †	257.02 *	257.99	256.75 †	261.84 *	262.28 *
WS, AP, APr	No NWP	259.43	257.44 †	261.22	261.10	263.39	268.49	263.59
WS, AP, APr	W62	249.26 *	249.21 * †	252.71 *	254.19 *	254.98 *	258.13 *	257.92 *
WS, AP, APr	W63	**243.82** *	246.35 *	249.99 *	251.27 *	253.84 *	255.92 *	256.67 *
WS, AP, APr	W64	247.38 *	248.97 *	252.73 *	254.89 *	255.60 *	260.46	260.15 *
WS, AP	No NWP	269.76	263.81 †	267.63	266.59	268.36	273.45	268.27
WS, AP	W62	255.81 *	253.52 * †	255.35 *	257.62 *	257.46 *	261.58 *	260.53 *
WS, AP	W63	248.44 *	250.92 *	253.47	255.76 *	256.22 *	261.10 *	259.65 *
WS, AP	W64	256.04 *	253.95 * †	258.66 *	261.17 *	260.14 *	265.55 *	264.33 *
WS, APr, GS, AP	No NWP	260.28	259.50 †	260.10	261.20	264.25	264.89	264.41
WS, APr, GS, AP	W62	252.25 *	250.55 * †	254.43 *	254.42 *	258.91 *	256.29 *	259.69 *
WS, APr, GS, AP	W63	247.06 *	248.17 *	250.56 *	254.25 *	256.56 *	255.82 *	259.42
WS, APr, GS, AP	W64	249.06 *	249.30 *	252.73 *	253.67 *	256.46	256.66 *	258.73 *
WS, APr, GS	No NWP	275.46	272.36 †	278.09	281.26	283.12	285.48	284.75
WS, APr, GS	W62	262.19 *	266.45 *	269.05 *	271.34 *	272.56 *	275.12 *	275.20 *
WS, APr, GS	W63	252.94 *	257.34	262.92	265.34 *	268.08 *	271.61 *	271.21 *
WS, APr, GS	W64	259.11 *	258.55 * †	266.17	267.02 *	269.95	276.42 *	275.67 *
WS	No NWP	299.02	293.60 †	299.05	296.26	298.94	299.84	299.95
WS	W62	274.06	278.05 *	281.11 *	280.75 *	279.34 *	281.74 *	281.85 *
WS	W63	270.57 *	269.14 * †	272.41 *	276.21	275.12 *	279.05 *	279.98 *
WS	W64	276.77 *	270.92 * †	276.50 *	277.68 *	281.50 *	280.79 *	285.75 *

Table A.16
MAE values (kW) for the **XGBoost** model with and without NWP. Values marked with † denote a statistically significant improvement (p-value<0.05) relative to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. Values marked with * indicate a significant improvement (p-value<0.05) compared to the No NWP counterpart. The p-values are calculated using the Diebold–Mariano test. Although using one-hour lagged variables yields to significant improvement, more lags does not lead to further improvement. The inclusion of any of the NWP variables leads to significant improvements.

Explanatory variables	NWP	0	1	2	3	4	6	12
AP, APr	No NWP	254.64	259.20	259.14	262.75	261.75	261.79	262.97
AP, APr	W62	246.65	244.11 * †	244.66	247.25 *	247.07	250.46 *	250.56 *
AP, APr	W63	**242.11** *	244.96 *	245.04	252.04 *	249.41	254.72 *	253.20 *
AP, APr	W64	245.70	245.79 *	247.78 *	248.43	249.68	252.52 *	253.75
AP	No NWP	256.61	255.44 †	256.52	259.35	257.75	258.18	262.76
AP	W62	248.60 *	244.55 * †	244.51 * †	248.86 *	246.37 *	250.43 *	249.78 *
AP	W63	245.16 *	246.62 *	247.86 *	251.13 *	249.58 *	255.27 *	253.31 *
AP	W64	249.24 *	248.90 * †	249.13 *	251.65 *	250.76 *	251.06 *	254.95
WS, AP, APr	No NWP	261.61	261.58 †	265.42	264.81	265.74	265.42	267.16
WS, AP, APr	W62	250.63 *	249.86 * †	250.44 *	251.57 *	256.00 *	255.17 *	258.08 *
WS, AP, APr	W63	246.78 *	248.17 *	249.70 *	254.77 *	254.37 *	255.00 *	256.80 *
WS, AP, APr	W64	252.19 *	248.53 * †	253.18 *	254.56 *	254.13 *	258.06 *	255.03 *
WS, AP	No NWP	266.56	262.14 †	263.48	263.50	264.89	261.24 †	264.07
WS, AP	W62	255.50 *	248.11 *	253.21 * †	254.41 *	256.87 *	253.06 * †	255.66 *
WS, AP	W63	252.57 *	246.67 * †	252.67	258.11 *	255.65 *	259.22 *	258.57 *
WS, AP	W64	255.23 *	251.08 * †	252.01 *	257.88 *	252.34 *	257.17	259.03 *
WS, APr, GS, AP	No NWP	260.98	259.00 †	263.18	264.77	264.30	264.89	266.14
WS, APr, GS, AP	W62	251.32 *	242.97 * †	251.28 *	257.96 *	256.47 *	260.04 *	254.49 *
WS, APr, GS, AP	W63	247.70 *	244.73 * †	253.93 *	257.80 *	257.91 *	258.67 *	257.37 *
WS, APr, GS, AP	W64	253.44 *	247.31 * †	253.82	255.45 *	253.20 *	255.94 *	258.85 *
WS, APr, GS	No NWP	279.36	285.44	288.63	290.97	291.36	293.82	294.84
WS, APr, GS	W62	261.00 *	264.93 *	270.51 *	277.50 *	275.95 *	279.20 *	279.72 *
WS, APr, GS	W63	263.61	266.61 *	266.25 *	274.09 *	274.60 *	276.20 *	277.40
WS, APr, GS	W64	261.09 *	267.83 *	268.50 *	274.92	270.71 *	278.98 *	277.51 *
WS	No NWP	289.67	288.24 †	292.68	293.21	291.58	291.61	292.12
WS	W62	271.80 *	271.62 * †	274.54 *	277.80 *	274.08 *	278.30 *	278.22 *
WS	W63	264.50 *	269.69 *	272.97 *	275.95 *	275.82 *	280.50 *	278.71 *
WS	W64	272.59 *	273.67 *	280.09 *	279.03 *	279.97 *	282.44 *	282.26 *

Table A.17
MAE values (kW) for the **LightGBM** model with and without NWP. Values marked with † denote a statistically significant improvement (p-value<0.05) relative to the last model with fewer lags that achieved significant improvement; if no intermediate lagged model is significant, the comparison is made against the 0-lag specification. Values marked with * indicate a significant improvement (p-value<0.05) compared to the No NWP counterpart. The p-values are calculated using the Diebold–Mariano test. Although using one-hour lagged variables yields to significant improvement, more lags does not lead to further improvement. The inclusion of any of the NWP variables leads to significant improvements.

Explanatory variables	NWP	0	1	2	3	4	6	12
AP, APr	No NWP	254.58	254.86	257.23	271.05	266.80	260.09	262.81
AP, APr	W62	246.88 *	236.32	248.87	245.82	249.08 *	251.98 *	252.04 *
AP, APr	W63	246.53 *	**236.25** * †	249.79	250.27 *	250.26	257.61	256.99 *
AP, APr	W64	246.19	244.60 * †	246.77	246.16	246.77	252.98	252.23 *
AP	No NWP	255.98	256.31	260.00	257.95	267.24	268.27	261.94
AP	W62	248.87 *	247.01 * †	247.99 *	250.98 *	251.26 *	251.30 *	253.51 *
AP	W63	247.34 *	244.29 * †	246.14 *	252.50 *	262.71 *	259.69 *	257.43 *
AP	W64	250.05 *	250.84 *	250.04 * †	251.35 *	250.87 *	250.38 *	253.84
WS, AP, APr	No NWP	255.29	266.90	262.40	264.99	265.49	264.70	269.14
WS, AP, APr	W62	247.54 *	246.27 * †	245.89 * †	251.66 *	251.14 *	256.92 *	259.81 *
WS, AP, APr	W63	243.45 *	250.04 *	252.73 *	254.80	251.87 *	257.42 *	258.89 *
WS, AP, APr	W64	241.97 *	244.55 *	249.90 *	249.17 *	254.98 *	253.46 *	256.40 *
WS, AP	No NWP	264.86	265.45	265.66	279.78	264.05 †	265.44	268.66
WS, AP	W62	250.50 *	254.42 *	248.18 * †	268.08 *	255.84 *	258.84 *	259.05 *
WS, AP	W63	243.08 *	251.64 *	251.58 *	256.80 *	255.46 *	259.06 *	258.65 *
WS, AP	W64	253.61 *	247.19 *	247.64 * †	259.07 *	252.58 *	268.60	258.75 *
WS, APr, GS, AP	No NWP	260.57	259.57 †	266.18	271.81	264.21	267.52	268.64
WS, APr, GS, AP	W62	247.51 *	242.57 * †	251.82 *	250.53 *	248.59 *	254.52 *	260.09 *
WS, APr, GS, AP	W63	241.10 *	240.98 †	247.69 *	256.73 *	258.83	255.96 *	257.24 *
WS, APr, GS, AP	W64	253.70 *	242.79 * †	249.41 *	251.32 *	252.19 *	256.05 *	258.83 *
WS, APr, GS	No NWP	289.32	288.98 †	291.03	292.42	295.36	296.07	295.98
WS, APr, GS	W62	265.43	266.08 *	268.54 *	275.83 *	276.21 *	273.72 *	277.50 *
WS, APr, GS	W63	258.91	259.71 *	267.30 *	272.86 *	274.13 *	277.00 *	279.20 *
WS, APr, GS	W64	270.08 *	258.46 * †	269.57 *	272.54 *	275.86 *	269.07 *	276.91 *
WS	No NWP	290.67	290.99	300.33	293.71	293.52	293.93	304.73
WS	W62	279.42 *	280.17 *	280.27 *	278.36 * †	276.68 * †	284.38 *	280.12 *
WS	W63	270.73 *	270.59 * †	279.51 *	279.95 *	275.61 *	278.51 *	284.47 *
WS	W64	272.04	269.97 * †	275.07 *	277.91 *	281.09 *	276.71 *	285.57 *

Data availability

The datasets used in this study are not publicly available due to distribution restrictions. The Alpha Ventus offshore wind farm data can be accessed through the RAVE platform. [14] Numerical Weather Prediction (NWP) variables from the ICON-D2 model are available via the Pamore database. [41] Access to both datasets requires compliance with their respective data use policies and licensing agreements.

References

[1] Desalegn Belachew, Gebeyehu Desta, Tamrat Bimrew, Tadiwose Tassew, Lata Ababa. Onshore versus offshore wind power trends and recent study practices in modeling of wind turbines' life-cycle impact assessments. Clean Eng Technol 2023;17:100691.

[2] Foley Aoife M, Leahy Paul G, Marvuglia Antonino, McKeogh Eamon J. Current methods and advances in forecasting of wind power generation. Renew Energy 2012;37(1):1–8.

[3] Zheng Chong Wei, Li Chong Yin, Pan Jing, Liu Ming Yang, Xia Lin Lin. An overview of global ocean wind energy resource evaluations. Renew Sustain Energy Rev 2016;53:1240–51.

[4] Hong Sunghun, McMorland Jade, Zhang Houxiang, Collu Maurizio, Halse Karl Henning. Floating offshore wind farm installation, challenges and opportunities: A comprehensive survey. Ocean Eng 2024;304:117793.

[5] Arslan Tuncar Ezgi, Sağlam Şafak, Oral Bülent. A review of short-term wind power generation forecasting methods in recent technological trends. Energy Rep 2024;12:197–209.

[6] Barthelmie RJ, Murray F, Pryor SC. The economic benefit of short-term forecasting for wind energy in the UK electricity market. Energy Policy 2008;36(5):1687–96.

[7] Qureshi Shezeena, Shaikh Faheemullah, Kumar Laveet, Ali Farooque, Awais Muhammad, Gürel Ali Etem. Short-term forecasting of wind power generation using artificial intelligence. Environ Challenges 2023;11:100722.

[8] Sun Yanmiao, Han Weixue. A review of enhancing wind power with AI: applications, economic implications, and green innovations. Digital Econ Sustain Development 2025;3(1):11.

[9] Das Prangon, Mashiata Maisha, Iglesias Gregorio. Big data meets big wind: A scientometric review of machine learning approaches in offshore wind energy. Energy AI 2024;18:100418.

[10] Alkhayat Ghadah, Mehmood Rashid. A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 2021;4:100060.

[11] Koshkarbay Nursultan, Mekhilef Saad, Saymbetov Ahmet, Kuttybay Nurzhigit, Nurgaliyev Madiyar, Dosymbetova Gulbakhar, Orynbassar Sayat, Yershov Evan, Kapparova Ainur, Zholamanov Batyrbek, et al. Adaptive control systems for dual axis tracker using clear sky index and output power forecasting based on ML in overcast weather conditions. Energy AI 2024;18:100432.

[12] Dhungana Hariom. A machine learning approach for wind turbine power forecasting for maintenance planning. Energy Inform 2025;8(1):2.

[13] Mo Site, Wang Haoxin, Li Bixiong, Xue Zhe, Fan Songhai, Liu Xianggen. Powerformer: A temporal-based transformer model for wind power forecasting. Energy Rep 2024;11:736–44.

[14] Fraunhofer IWES. RAVE Offshore research at alpha ventus. 2009.

[15] Federal Maritime and Hydrographic Agency. Quality control of RAVE measurements from AV00, AV04, AV05, AV07–AV12 and FINO1. 2019.

[16] Tuncar Ezgi Arslan, Sağlam Şafak, Oral Bülent. A review of short-term wind power generation forecasting methods in recent technological trends. Energy Rep 2024;12:197–209.

[17] Waqas Muhammad, Humphries Usa Wannasingha, Chueasa Bunthid, Wangwongchai Angkool. Artificial intelligence and numerical weather prediction models: A technical survey. Nat Hazards Res 2024.

[18] Alizadeh Omid. Advances and challenges in climate modeling. Climatic Change 2022;170(1):18.

[19] Van Poecke Aaron, Tabari Hossein, Hellinckx Peter. Unveiling the backbone of the renewable energy forecasting process: Exploring direct and indirect methods and their applications. Energy Rep 2024;11:544–57.

[20] Makridakis Spyros, Spiliotis Evangelos, Assimakopoulos Vassilios. Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS One 2018;13(3):1–26.

[21] Hanifi Shahram, Zare-Behtash Hossein, Cammarano Andrea, Lotfian Saeid. Offshore wind power forecasting based on WPD and optimised deep learning methods. Renew Energy 2023;218:119241.

[22] Torres JL, García A, De Blas M, De Francisco A. Forecast of hourly average wind speed with ARMA models in navarre (Spain). Sol Energy 2005;79(1):65–77.

[23] Erdem Ergin, Shi Jing. ARMA based approaches for forecasting the tuple of wind speed and direction. Appl Energy 2011;88(4):1405–14.

[24] Neshat Mehdi, Nezhad Meysam Majidi, Abbasnejad Ehsan, Mirjalili Seyedali, Tjernberg Lina Bertling, Astiaso Garcia Davide, Alexander Bradley, Wagner Markus. A deep learning-based evolutionary model for short-term wind speed forecasting: A case study of the lillgrund offshore wind farm. Energy Convers Manage 2021;236:114002.

[25] Huang Cong, Karimi Hamid Reza, Mei Peng, Yang Daoguang, Shi Quan. Evolving long short-term memory neural network for wind speed forecasting. Inform Sci 2023;632:390–410.

[26] Wan Shanshan, Yang Lan, Ding Keliang, Qiu Dongwei. Dynamic gesture recognition based on three-stream coordinate attention network and knowledge distillation. IEEE Access 2023;11:50547–59.

[27] Su Yongxin, Yu Jing, Tan Mao, Wu Zexuan, Xiao Zhe, Hu Jianghui. A LSTM based wind power forecasting method considering wind frequency components and the wind turbine states. In: 2019 22nd international conference on electrical machines and systems. 2019, p. 1–6.

[28] Yin Xiuxing, Zhao Xiaowei. Big data driven multi-objective predictions for offshore wind farm based on machine learning algorithms. Energy 2019;186:115704.

[29] Hanifi Shahram, Lotfian Saeid, Zare-Behtash Hossein, Cammarano Andrea. Offshore wind power forecasting—A new hyperparameter optimisation algorithm for deep learning models. Energies 2022;15(19).

[30] Ko Min-Seung, Lee Kwangsuk, Kim Jae-Kyeong, Hong Chang Woo, Dong Zhao Yang, Hur Kyeon. Deep concatenated residual network with bidirectional LSTM for one-hour-ahead wind power forecasting. IEEE Trans Sustain Energy 2021;12(2):1321–35.

[31] Zu XR, Song RX. Short-term wind power prediction method based on wavelet packet decomposition and improved GRU. J Phys: Conf Ser 2018;1087(2):022034.

[32] Su Yongxin, Yu Jing, Tan Mao, Wu Zexuan, Xiao Zhe, Hu Jianghui. A LSTM based wind power forecasting method considering wind frequency components and the wind turbine states. In: 2019 22nd international conference on electrical machines and systems. 2019, p. 1–6.

[33] Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory. Neural Comput 1997;9(8):1735–80.

[34] Alkanhel Reem, El-kenawy El-Sayed M, Abdelhamid Abdelaziz A, Ibrahim Abdelhameed, Alohali Manal Abdullah, Abotaleb Mostafa, Khafaga Doaa Sami. Network intrusion detection based on feature selection and hybrid metaheuristic optimization. Comput Mater Contin 2022;74(2):2677–93.

[35] El-Kenawy El-Sayed M, Khodadadi Nima, Mirjalili Seyedali, Makarovskikh Tatiana, Abotaleb Mostafa, Karim Faten Khalid, Alkahtani Hend K, Abdelhamid Abdelaziz A, Eid Marwa M, Horiuchi Takahiko, Ibrahim Abdelhameed, Khafaga Doaa Sami. Metaheuristic optimization for improving weed detection in wheat images captured by drones. Mathematics 2022;10(23).

[36] Atteia Ghada, El-kenawy El-Sayed M, Samee Nagwan Abdel, Jamjoom Mona M, Ibrahim Abdelhameed, Abdelhamid Abdelaziz A, Azar Ahmad Taher, Khodadadi Nima, Ghanem Reham A, Shams Mahmoud Y. Adaptive dynamic dipper throated optimization for feature selection in medical data. Comput Mater Contin 2023;75(1):1883–900.

[37] Myriam Hadjouni, A. Abdelhamid Abdelaziz, El-Kenawy El-Sayed M, Ibrahim Abdelhameed, Eid Marwa Metwally, Jamjoom Mona M, Khafaga Doaa Sami. Advanced meta-heuristic algorithm based on particle swarm and al-biruni earth radius optimization methods for oral cancer detection. IEEE Access 2023;11:23681–700.

[38] Akiba Takuya, Sano Shotaro, Yanase Toshihiko, Ohta Takeru, Koyama Masanori. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2623–31.

[39] Rajaperumal TA, Christopher Columbus C. Enhanced wind power forecasting using machine learning, deep learning models and ensemble integration. Sci Rep 2025;15(1):20572.

[40] Bentéjac Candice, Csörgő Anna, Martínez-Muñoz Gonzalo. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54(3):1937–67.

[41] Deutscher Wetterdienst. Pamore - retrieving archived forecast model data. 2025, https://www.dwd.de/EN/ourservices/pamore/pamore.html. (Accessed 24 May 2025).

[42] Christian Sebastian Lamprecht. Meteostat python. 2024.

[43] Diebold Francis X, Mariano Robert S. Comparing predictive accuracy. J Bus Econom Statist 2002;20(1):134–44.

[44] Lundberg Scott M, Lee Su-In. A unified approach to interpreting model predictions. Curran Associates, Inc.; 2017.