AIMS Energy

AIMS Energy, 14(2): 387–417.
DOI: 10.3934/energy.2026017
Received: 25 August 2025
Revised: 16 February 2026
Accepted: 06 March 2026
Published: 24 March 2026

https://www.aimspress.com/journal/energy

Research article

DVTransformer: A novel approach for multi-step wind power forecasting using spatio-temporal dynamics and predictive strategies

Syed Muhammad Rashid Hussain[1], Mirza Muhammad Ali Baig[1] and Muhammad Uzair Yousuf[2,*]

[1] Department of Electrical Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan
[2] Department of Mechanical Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan

* **Correspondence:** Email: uzairyousuf@neduet.edu.pk; Tel: +922199261261.

Abstract: Short-term wind power forecasting is essential for wind farm management and reliable grid operations. However, the accuracy of turbine-specific forecasting is often compromised by limitations in sequence modeling and misleading information from the surrounding wind turbines. To address these challenges, we proposed a novel DVTransformer (DTW-VARIMA-Transformer) framework for turbine specific forecast integrating Transformer neural networks and a predictive strategy. This approach integrates spatio-temporal dynamics using wind speed from the surrounding turbines to forecast the wind power of the target turbine in a wind farm. Wind turbines were selected using dynamic time warping (DTW) based metrics, which calculate the dynamic distance between wind speed time series, ensuring the reliability of spatial information. Additionally, we incorporated predicted wind speeds of surrounding turbines using vector autoregressive integrated moving average (VARIMA), alongside historical data, to better capture the influence of future wind conditions on the target turbine power output. The DVTransformer performance was assessed against parallel models under various metrics, including mean absolute error (MAE), mean squared error (MSE), and correlation coefficient (R), demonstrating significant improvements in a multi-step ahead forecasting task. The proposed DVTransformer was first validated on two representative turbines to assess turbine-specific forecasting performance. For 3-step-ahead forecasting, DVTransformer achieved an average MSE of 0.085 for turbine T01, corresponding to reductions of 34.92%, 24.77%, 2.20%, and 27.34% compared to the Fast Fourier Transformer (FFTransformer), Informer, Spatio-temporal Long Short-Term Memory (ST-LSTM), and Spatio-temporal Multi-Layer Perceptron (ST-MLP), respectively. Similarly, for turbine T05, the proposed model attained an average MSE of 0.090,

achieving MSE reductions of 38.45%, 23.08%, 2.93%, and 25.96% against the same benchmark models, demonstrating robustness across turbines. To further evaluate computational efficiency and scalability, a training-budget sensitivity analysis was conducted by comparing a DVTransformer trained for a single epoch against a fully trained Transformer. The results showed that the DVTransformer achieved comparable prediction accuracy across all turbines while reducing significant computational time. Evident from incorporating reliable spatial information, employing predictive wind conditions and using Transformer to capture long and short-term dependencies within time sequences increased the overall performance of the proposed method.

Keywords: wind power forecasting; dynamic time warping (DTW); spatio-temporal correlation; projected wind conditions; Transformer neural network

1. Introduction

In recent years, wind power has become vital in the transition to low-carbon energy systems, driven by the global effort to mitigate climate change [1,2]. According to International Renewable Energy Agency (IRENA) 2024 statistics, global wind power capacity reached 1017.19 GW in 2023, and is continuing its steady growth [3]. However, despite its growing importance, the intermittent nature of wind and the unpredictability of local atmospheric conditions introduce volatility in power generation and complicate large-scale grid integration [4,5]. This variability poses significant challenges for maintaining grid stability and reliability [6,7]. To address these issues, the development of accurate and robust wind power forecasting techniques is essential for ensuring reliable integration of wind energy into the power grid [8–10].

Wind power forecasting models operate across time scales [11], including very short-term (a few seconds to 30 minutes ahead) [12], which assists in real-time turbine control and load tracking [13], short-term (30 minutes to 6 hours ahead) [14,15], which is used for load dispatch planning [16], medium-term (6 hours to 1 day ahead) [17] ,which supports energy trading and power system management, and long-term (1 day to 1 week or more ahead) [18] aids in optimizing maintenance scheduling [19,20]. Accurate short-term forecasts of wind turbine power outputs are particularly crucial to prevent power grid disruptions caused by abnormal fluctuations in energy sources [8,9].

Forecasting models can be divided into four categories: physical, statistical, machine learning and hybrid models [21]. Physical models forecast by considering meteorological factors (e.g., air pressure, humidity, and temperature), geographic information, but tend to be complex and less accurate for short-term forecasts [22]. On the other hand, statistical forecasting methods, such as the autoregressive integrated moving average (ARIMA) models [21], exponential smoothing [23,24], and grey model [25], rely on historical wind power data to identify potential correlations with future wind power. These statistical models are easier to build and generally offer higher accuracy for short-term forecasting compared to physical approaches, but their accuracy degrades as the prediction horizon or steps increase [22], which can be crucial for long-term forecasts.

Machine learning models like the artificial feed-forward neural network [26], Kalman filter [27], random forests [28], support vector machines [29], the fuzzy logic method [30], extreme learning machines [31], and the elman network [32] have been increasingly applied in wind power forecasting. These models offer improved accuracy by learning complex patterns in wind power generation. With the advancements in deep learning technology, deep neural networks (DNNs) are widely employed in

wind power forecasting due to their superior capability in dealing with complex nonlinear problems, which greatly enhance the time series feature learning capability [8]. Notably, long short-term memory (LSTM) networks [33] are favored for their ability to analyze time-series data and have been widely adapted for forecasting tasks. They have been further extended through hybrid neural network (HNN) models integrating convolutional neural network (CNN) for feature extraction [34]. Attention-based models, particularly Transformer neural networks [35], have risen to prominence in time series forecasting due to their superior performance for addressing the long-term dependencies in the time series data more effectively in recent years. Niu et al. [36] utilized an attention-based gated recurrent unit (GRU) network for wind power forecasting. Compared with the single forecasting model, integrating forecasting approaches constitute hybrid models with the goal to enhance prediction accuracy, though the hybridization does not always guarantee better performance [9].

In terms of forecasting objective, models are categorized into two types: Wind turbine forecasting, which predicts the power output of an individual turbine [37,38], and wind farm forecasting, which aggregates data from multiple turbines to predict the total output of the farm [39,40].

The operation of large-scale wind turbines necessitates forecasts with higher spatial resolution than traditional farm-level predictions. Turbine-level forecasting is particularly beneficial for improving operational strategies in complex wind farms, where the spatial variability of wind across turbines can significantly impact power generation and operational efficiency [41]. Moreover, precise turbine-level predictions can assist in turbine-specific operations such as power curve analysis [42] and estimation [43], structural health monitoring for turbines [44], damage-mitigating control [45], load alleviation [46], wake steering and turbine-to-plant power optimization [47], turbine degradation modeling [48], and maintenance scheduling [49]. In [41], turbine-specific short-term wind speed forecasting is performed considering within-farm wind field dependencies and fluctuations. Studies on single wind farm multi-turbine forecasting have demonstrated that aggregating turbine-level predictions can outperform direct farm-level models, as individual turbines experience heterogeneous wind conditions across large wind farms. Browell et al. [50] showed that wind farm power forecasts based on a conditional weighted aggregation of turbine-level predictions achieve superior accuracy compared to direct farm-level forecasting for horizons up to 48 hours. Similarly, Yakoub et al. [51] reported that aggregated turbine-level modeling improves wind power forecasting accuracy by approximately 10% and 15% in terms of root mean square error (RMSE) and mean absolute error (MAE), respectively, relative to conventional farm-level approaches. Ezzat [41] proposed a turbine-tailored probabilistic forecasting framework that couples spatio-temporal wind speed modeling with a power curve-based power conversion, achieving about 9% improvement over persistence forecasts and 7–9% over autoregressive and Gaussian process-based methods. Zhang et al. [37] demonstrated that LSTM-based wind turbine power forecasting, coupled with Gaussian mixture model (GMM) uncertainty modeling, can effectively support turbine-level power dispatching in modern grid operations. Deng et al. [52] proposed a turbine-level hybrid self attention-based deep auto regressive recurrent neural network (SA-DeepAR) model that corrects Supervisory Control and Data Acquisition (SCADA) wind speed with LSTM and uses self-attention to capture input correlations, achieving 44% improvement in short-term wind power prediction accuracy. Su et al. [53] proposed a spatio-temporal hybrid model combining CNN, bidirectional long short-term memory (BiLSTM), and graph convolutional networks (GCN) to capture turbine correlations and multivariate meteorological patterns while using Wasserstein Generative Adversarial Networks (WGAN) to handle missing data, significantly improving multi-turbine wind power forecasting accuracy. Furthermore, Sopeña et al. [54]

highlighted that high-resolution SCADA-based turbine-level wind speed and power measurements can significantly enhance short-term forecasting performance through turbine-tailored models; however, such fine-grained modeling incurs substantial computational overhead, motivating the need for approaches that balance forecasting accuracy and computational efficiency.

Spatial information plays a crucial role in wind power forecasting, as wind dynamics are inherently influenced by interactions among neighbouring turbines and wind farms. Understanding spatial dependencies among turbines can further refine forecasts and better account for spatial variability [7]. The spatial correlation between wind speed and direction was exploited in [55] by performing regression on spatial information conditioned on wind direction and in [56] through the application of vector autoregressive models. Dowell et al. [57] utilized spatial information by modeling the location parameter as a vector-valued spatio-temporal process, enabling the capture of spatial dependencies alongside temporal dynamics. Tastu et al. [58] demonstrated that leveraging multiple wind farms as spatial sensors significantly improves wind power forecasting accuracy at a target site. Yu et al. [59] developed a spatiotemporal wind speed prediction model based on graph attention networks (GAT) and GRUs to capture complex spatial and temporal dependencies. Zhen et al. [60] proposed BiLSTM-CNN considering temporal-spatial feature extraction for wind power forecasting. Yu et al. [61] introduced a regional wind power prediction approach employing spatiotemporal clustering and a hybrid neural network to effectively learn latent spatial-temporal dependencies among wind farms.

Despite the progress in wind power forecasting, some gaps remain unaddressed. First, there is limited exploration of integrating predicted wind speeds alongside historical data to enhance turbine-specific forecasts. Second, models often aggregate wind power predictions across farms, overlooking the unique spatial characteristics and interactions between individual turbines. Moreover, while studies emphasize sophisticated deep learning models, simpler yet effective architectures can be equally competitive if spatio-temporal information is effectively utilized [62,63].

To address these gaps, we propose a novel DVTransformer approach that utilizes temporal and spatial features between wind turbines along with the projected wind conditions to enhance forecasting accuracy. This is achieved by analyzing the spatial distribution, dependence of data in similar wind conditions, and dynamic contextual information between the wind turbines to elevate wind power forecasting. Instead of aggregating predictions across wind farms, the proposed DVTransformer approach focuses on individual turbines using wind speed and wind power data to model spatial correlations. Spatial information reliability is ensured using Dynamic Time Warping (DTW). Incorporating predicted wind speeds from neighboring turbines, alongside historical data, enhances wind power predictions for a target turbine.

The contributions of this paper can be summarized as follows:

1. **Integration of projected wind conditions in a spatio-temporal forecasting framework:** The proposed method utilizes a novel approach of wind speed predictor to project the prevailing wind conditions of the surrounding turbines and utilize it in spatio-temporal transformer attention architecture to forecast the wind power of a target turbine. By integrating these predicted values, the model can better anticipate future wind conditions for the target turbine and its impact on power production, leading to more precise forecasting outcomes.
2. **Spatial information is analyzed to capture long-term dynamic spatial characteristics:** DTWis used to evaluate the similarity of wind-speed patterns among turbines, ensuring that only

relevant spatial information contributes to forecasting. This reduces interference from unrelated or noisy spatial data, enhancing prediction reliability.

3. **Efficiency-focused forecasting with minimal training**: The proposed DVTransformer achieves competitive multi-step wind power predictions while training for only a single epoch, demonstrating that high accuracy can be achieved under constrained computational budgets, which highlights the framework's efficiency and practicality.

The effectiveness of the DVTransformer is validated through multi-step wind power predictions on a real-world dataset, demonstrating its competitive performance against state-of-the-art models in comparative analyses.

The rest of this paper is structured as follows: In Section 2, we provide an overview of the proposed methodology, along with a comprehensive description of the techniques employed. In Section 3, the datasets used in the study are presented. In Section 4, we briefly describe the evaluation metrics and experimental setup. In Section 5, we deliver a thorough analysis of the experimental findings and discuss the performance comparison between state-of-the-art forecasting models. In Section 6, we summarize the key conclusions of this work and suggest avenues for future research in Section 7.

2. Framework of the proposed method

In this section, we provide fundamental details on which the proposed novel method for short-term multi-step wind power forecasting, using spatio-temporal information and predicted wind conditions based on Transformer neural network, is formulated.

To enhance the accuracy of spatial information and alleviate the risks associated with inadequate long-term dependency modeling, a novel method for multi-step wind power forecasting is proposed, illustrated in schematic Figure 1. Historical wind speed and power data are collected from wind turbines and undergo pre-processing steps to normalize amplitude and address any missing data issues. To enhance the reliability of wind condition assessments, spatial information is evaluated using DTW. This approach aims to improve forecasting accuracy by integrating valid measurements from neighboring wind turbines and employing predictive strategies.

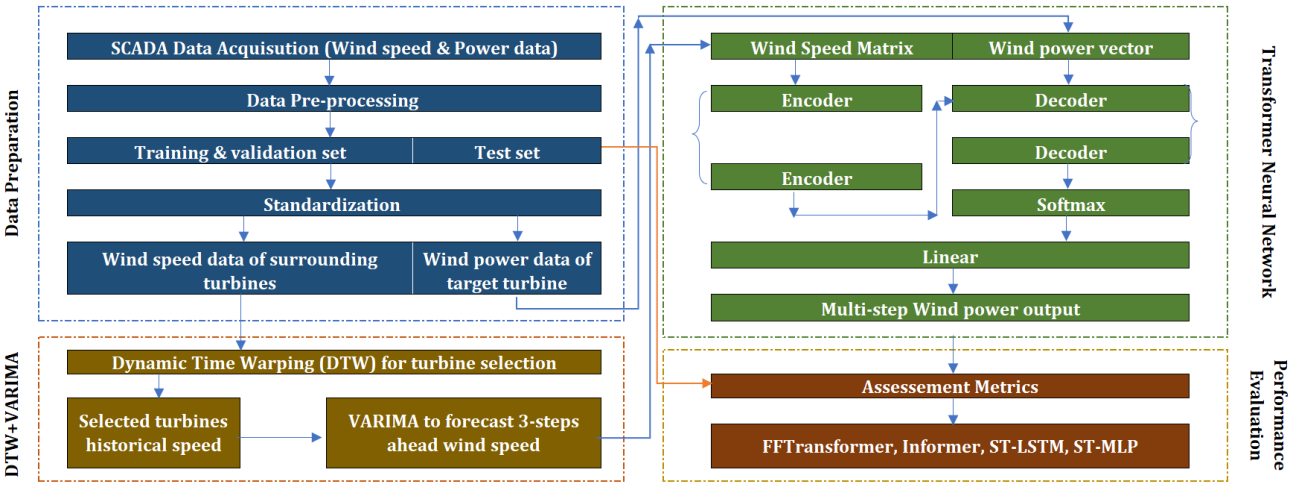

Figure 1. Framework of the proposed DVTransformer.

Additionally, Transformer-based neural networks [64] are utilized to effectively capture the intrinsic relationships between wind conditions and power outputs. Specifically, latent features from the wind speeds of surrounding turbines are extracted via encoders and then processed through cascaded decoders. The encoders and decoders employ a multi-head attention mechanism to model dependencies within sequences, regardless of distance, thereby addressing the challenge of long-term dependencies. Once the neural networks are developed and trained, real-time multi-step wind power predictions can be conducted using wind condition observations and power output data. The details of the proposed method are elaborated in the following subsections.

2.1. Data acquisition & pre-processing

The collection of wind power and wind speed data is done through SCADA and is typically averaged over ten-minute intervals to support short-term forecasting. With industrial internet of things (IIoT) advancements, this data is transmitted to central systems [65], but it may contain anomalies due to technical and environmental issues such as an absence of wind during run-up, anemometer defect, icing on anemometer, frequency converter error, tower resonance, sensor error, manual brake, and safety stop. It is mandatory to remove these anomalies by applying data pre-processing techniques.

Three variations of anomalies are reflected in the dataset in terms of (i) missing values (ii) outliers and (iii) negative power values. Missing values are usually represented in the form of NaN values. With relatively low proportion of missing entries, the affected instances are discarded. However, when the proportion is higher, it becomes necessary to identify the nature of the data loss and apply advanced data correction techniques. In this study, the missing data constitute only a small percentage of the total measurements and, therefore, these instances are removed. Furthermore, the Inter-Quartile Regression (IQR) method [66] is applied to the data, which enhances the accuracy of the dataset statistics by dropping outlying points. Moreover, negative power values are also removed from the dataset. The dataset is further standardized by removing the mean and scaling to unit variance for each feature, as given in Eq (1), where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\sigma}$ is the standard deviation of the training samples.

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

From the pre-processed data, selection criteria of surrounding turbines for a particular target turbine wind power forecast are discussed in the subsequent subsection.

2.2. Reliability of spatial information using DTW

Wind generation is driven by the movement of air masses from high-pressure to low-pressure regions, resulting in similar wind patterns at various turbines within a farm due to meteorological inertia [67,68]. For example, if the wind is blowing southwest, the wind at turbine T6 will reach T4 and T5, leading to strong correlations in wind speed among these turbines, as illustrated in Figure 2. However, real-world environmental factors like terrain, surface roughness, and vegetation can alter airflow, causing variability in wind conditions at the target turbine [69]. Such uncorrelated wind data may introduce noise into the spatial information, and it is difficult to directly assess the significance

of each wind turbine in forecasting performance. Therefore, it is essential to assess the correlations between the target wind turbine and surrounding turbines.

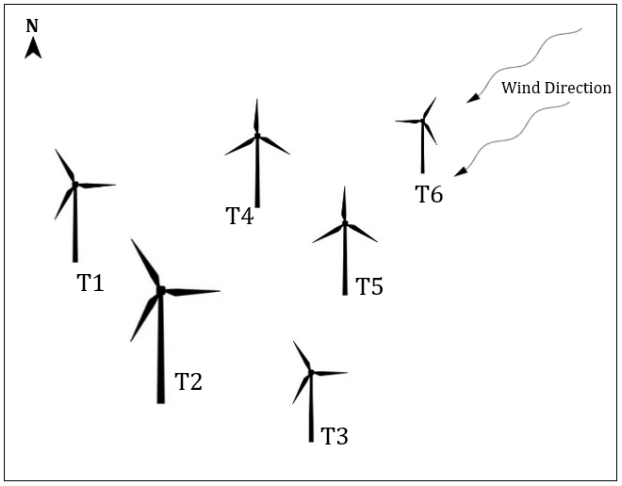

Figure 2. Example layout of a wind farm with multiple turbines.

2.2.1. Dynamic Time Warping (DTW)

In the proposed method, DTW is utilized to evaluate these correlations and serves as an indicator to determine the relative significance of the surrounding wind turbines with the target turbine. DTW is an algorithm that measures the similarity between two time series by aligning them in a non-linear fashion, enabling time shifts and distortions [70]. DTW is therefore employed to measure the similarity of dynamic characteristics, using the wind speed time series, of surrounding wind turbines.

For two wind speed time series $\boldsymbol{X} = \{x_1, x_2, \ldots, x_m\}$ and $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$, where $(m \leq n)$, DTW computes the optimal alignment by evaluating the distance $d(i \leq j) = |x_i, y_i|^P$ between their elements, with $|.|^P$ representing the P norm. A distance matrix D_{mxn} is constructed, with each element representing the distance between corresponding points in $\boldsymbol{X}$ and $\boldsymbol{Y}$. The DTW then identifies a warping path $\boldsymbol{W}_{DTW}$ as represented by Eq (2), defines the optimal alignment sequence by minimizing the cumulative distance metric $d(i, j)$. The final distance reflects the accumulated cost of the optimal alignment path.

$$\boldsymbol{W}_{DTW} = \{w_1, w_2, \ldots, w_k\}, \ \ (max(m, n) \leq k \ \leq m + n \ - 1) \tag{2}$$

The warping path must adhere to the following constraints:

i. Boundary: $w_1 = (1,1), \ w_k = (m, n)$
ii. Continuity: Path can move only to adjacent points
iii. Monotonicity: Adjacent points on the warping path must satisfy:

$$w_{i+1} - \ w_1 \in \{(1,0), (0,1), (1,1)\}, \ \ (l = 1,2, \ldots, k - 1)$$

The DTW uses dynamic programming to find the optimal path known as the warped path that minimizes cumulative costs through the cost matrix, aligning sequences by moving right, up, or diagonally, as shown by Eq (3).

$$DTW(i,j) = d(i,j) + min\begin{cases} DTW(i-1,j) \\ DTW(i,j-1) \\ DTW(i-1,j-1) \end{cases} \quad (3)$$

DTW evaluates the distance between the wind speeds of surrounding turbines and the least metrics score, which reveals the maximum correlation of wind conditions between wind turbines. To mitigate the interference from uncorrelated data of the surrounding wind turbines, adjacent neighboring wind turbines with minimum DTW scores are selected for the target wind turbine. By integrating data from these shortlisted wind turbines, we can gain a more comprehensive understanding of wind conditions compared to using single-turbine measurements.

2.3. Modeling projected wind conditions

Consider a farm with multiple wind turbines, where the objective is to predict the power output of a specific target turbine for instance T2 as in Figure 2. The target turbine T2 is influenced by wind conditions from multiple upstream or lateral turbines, such as T4, T5, and T6, and other surrounding turbines. Therefore, the wind speed is predicted for the upstream or correlated turbines, as it is valuable for the target turbine power prediction, since wind energy is highly correlated to wind speed at the wind site [71]. The future predicted wind data of these turbines supplement the wind power prediction of target turbine.

Therefore, a predictor is employed to predict the wind speed for the turbines shortlisted by DTW to understand local wind field and its evolution in a broader perspective with the help of historical wind conditions. By modeling this wind field evolution, the proposed method can better account for wind flow patterns and their collective impact on the power output of the target turbine.

2.3.1. Vector Autoregressive Integrated Moving Average (VARIMA)

This predictor is modeled by applying the VARIMA model. It extends the traditional ARIMA [72] framework to support multivariate time series data by capturing the dynamics between them, addressing limitations of univariate analysis where significant external variables are overlooked [73]. Furthermore, this multivariate approach enables a more comprehensive representation of variable relationships, leading to more robust predictions [74,75]. Mathematically, it can be expressed as in Eq (4):

$$\boldsymbol{\Phi}(B)(1-B)^d \boldsymbol{y_t} = \boldsymbol{\delta} + \boldsymbol{\Theta}(B)\boldsymbol{\varepsilon_t} \quad (4)$$

where
$\boldsymbol{y_t}$: n variables at t time − $(n\ x\ 1)$,
$\boldsymbol{\Phi}$: coefficient matrix of Vector Autoregressive (VAR) − $(n\ x\ n)$,
$\boldsymbol{\delta}$: vector averages $(n\ x\ 1)$,
$\boldsymbol{\Theta}$: coefficient matrix of Vector Moving Average (VMA) − $(n\ x\ n)$,
$(1-B)^d$: differencing components,
$\boldsymbol{\varepsilon_t}$: error vector $(n\ x\ 1)$,
B: backshift operator.

VARIMA modeling generally follows three major steps. First, model identification is performed by examining the stationarity of the multivariate time series and determining the appropriate differencing order (d). The autoregressive order (p) and moving-average order (q) are then selected

using autocorrelation analysis in conjunction with information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Second, parameter estimation is carried out to jointly learn the coefficients of the vector autoregressive and moving-average components across all variables, typically using maximum likelihood or least-squares methods, enabling the model to capture temporal dynamics and cross-variable dependencies. Finally, diagnostic checking is conducted through residual analysis to ensure independence and white-noise behavior, after which the validated VARIMA (*p, d, q*) model is employed for multi-step-ahead forecasting.

Based on the assumption that precise future wind speed data enhances power prediction accuracy, a VARIMA (*p, d, q*) model is employed as a "wind speed predictor" for surrounding turbines, given that true future wind speeds are unknown in advance. The model is fitted using historical wind speed data from the selected neighboring turbines. The optimal model orders (*p, d, q*) are determined using AIC and BIC. A rolling-window strategy with a window size of 32-time stamps is employed to update the model periodically with the latest observations. The window advances with a slide of a one-time step, such that each new input sequence is formed by discarding the oldest observation and appending the most recent available value, as shown in Figure 3(a). To prevent data leakage, the VARIMA model is trained using only historical observations available to each prediction time point. Forecasts are generated in a strictly forward-chaining manner, ensuring that no future information is used during model estimation. Multi-step-ahead wind speed forecasts are subsequently generated in a recursive manner using the fitted model. Specifically, for each input sequence length, the VARIMA model forecasts wind speeds 3-steps ahead for DTW based selected turbines. The sequence is then updated by embedding these predicted values at the end and discarding the initial three wind speed values. For instance, historical wind speed sequence data from 00:30 to 05:10, including VARIMA predicted wind speeds from 05:20 to 05:40, is used to predict the wind power output from 05:20 to 05:40 using DVTransformer with a temporal resolution of 10 mins, as shown in Figure 3(b).

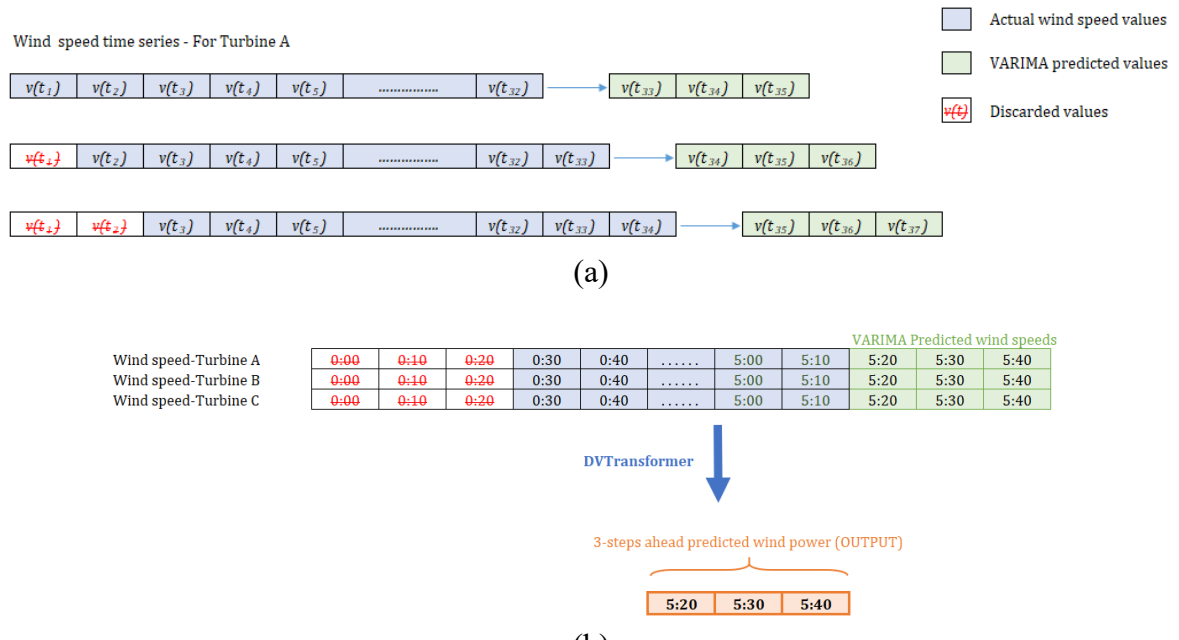

Figure 3. (a) Rolling mechanism with a window size of 32; and (b) Historical and VARIMA predicted wind speeds for multi-step wind power forecasting using DVTransformer.

2.4. DVTransformer for wind power modeling

The objective of DVTransformer is to estimate the relationship between surrounding wind behaviors and power production. The updated wind speed matrix ($\boldsymbol{W_s}$), as shown in Eq 5, serves as the encoder input for wind power modeling, containing the historical and predicted wind speeds of three surrounding wind turbines ($\boldsymbol{w_{s,T_A}}, \boldsymbol{w_{s,T_B}}, \boldsymbol{w_{s,T_C}}$) along with a date-time feature ($\boldsymbol{x_{dt}}$). As wind conditions are highly sensitive to daily weather variations, the date-time feature for the target turbine is derived from the sampling intervals of SCADA data. The updated sequence is then fed into the encoder, which extracts relevant information from this sequence for accurate power forecasting. The process is repeated for every sequence length during the training and inference of the Transformer neural networks.

$$\boldsymbol{W_s} = \left\{\boldsymbol{x_{dt}}, \boldsymbol{w_{s,T_A}}, \boldsymbol{w_{s,T_B}}, \boldsymbol{w_{s,T_C}}\right\} \tag{5}$$

In the decoder stage, current wind power measurements are also employed to represent the initial state of sequences. The output of encoder-decoder Transformer neural network is the power vector $\boldsymbol{p_L}$, as represented in Eq (6), containing the multi-step ahead wind power of the target turbine that is intended to be forecasted, and $\boldsymbol{L} = \{\boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3}, \ldots, \boldsymbol{L}\} \in \mathbb{R}$ is the forecasting horizon.

$$\boldsymbol{p_L} = \{p_1, p_2, p_3, \ldots, p_L\} \in \mathbb{R} \tag{6}$$

The conversion from wind conditions to wind power is achieved by the transformer employing N number of encoders ($N_{encoders}$) and decoders ($N_{decoders}$). The encoder module is a fundamental component of Transformer-based architectures, consisting of an input embedding layer, a positional encoding layer, and encoding layers. As shown in Figure 4, the input historical wind conditions undergo several transformations. Initially, the sequence is projected into a higher-dimensional space through the input embedding layer. At each time step t, the historical values of the wind matrix are mapped to d-dimensional vectors via a fully connected layer. To preserve the temporal order of the sequence, the positional encoding PE module is employed, which encodes positional information using sin and cos functions as defined in Eqs (7) & (8). In this formulation, pos refers to the position within the sequence, i indicates the dimensional index of the embedded vector, and d_{model} represents the embedding dimension. The position encoding is then combined with the input embeddings, which are subsequently passed to the encoder.

$$PE\ (pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{7}$$

$$PE\ (pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{8}$$

Each encoder layer contains a multi-head attention mechanism and a feedforward neural network. The multi-head attention mechanism, a critical component of Transformer neural networks, facilitates the flexible mapping between input and output sequences through a process analogous to dictionary look-up. By employing multiple attention heads, the model can simultaneously focus on various segments of the input sequence, enabling each head to capture unique temporal or contextual

relationships. This enhances the model's ability to learn complex dependencies, improving prediction performance and robustness. The multi-head attention mechanism applies three projection matrices, $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, $\boldsymbol{W_V}$, to the input vectors, generating the query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ matrices, as described in Eq (9), which are subsequently utilized in the decoder module.

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W_{s,Q}} P_{input\_embedding,} \\ \boldsymbol{K} = \boldsymbol{W_{s,K}} P_{input\_embedding,} \\ \boldsymbol{V} = \boldsymbol{W_{s,V}} P_{input\_embedding.} \end{cases} \tag{9}$$

The attention weights are computed using a $softamx$ function, as shown in Eq (10), which determines the relevance of each input with the output prediction, where $d_k = d_{model}$.

$$attention = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

The $softmax$ function plays a crucial role by transforming the raw attention scores into a normalized probability distribution that reflects the relative importance of each input element with respect to the current query. Specifically, it amplifies larger similarity scores while suppressing less relevant ones, ensuring that highly correlated input positions receive greater emphasis in the attention output. This normalization enables stable training and enables the model to selectively aggregate information from the most relevant temporal contexts.

Mathematically, the $softmax$ function converts a vector of real numbers into a probability distribution, with each element rescaled to lie between 0 and 1, and the total sum of elements equal to 1, where x_i is the i-th element of the input vector, and the denominator ensures normalization across all elements.

$$softmax\,(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{11}$$

Through this mechanism, the Transformer effectively assigns adaptive weights to time steps, enabling dynamic focus on the most informative parts of the input sequence during prediction.

The decoders, which are interconnected in a cascade, utilize these latent features alongside historical data to produce the predictions. This methodology enables the model to effectively combine the temporal and spatial patterns learned by the encoders with the historical wind data, thereby enhancing the accuracy and reliability of the wind power forecasts.

In the encoders and decoders, a multi-layer perceptron (MLP) is incorporated to improve the model capability. Layer normalization is employed to mitigate the impact of varying data magnitudes on wind power modeling, ensuring more stable and accurate predictions. Moreover, using the developed DVTransformer neural network, the process of wind power modeling is facilitated through the transformation of sequential wind condition data into forecasts of future power outputs.

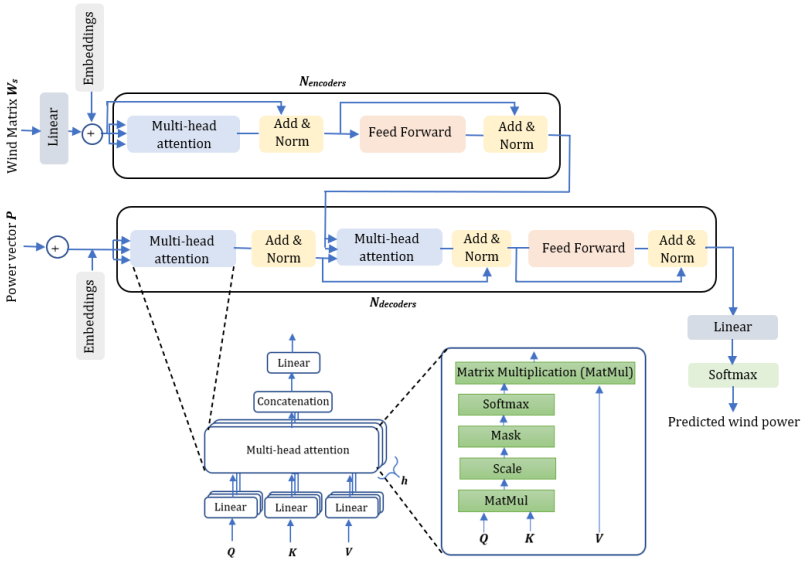

Figure 4. Wind power modeling with the DVTransformer neural network.

3. Data description

In this study, Penmanshiel wind farm is selected. It is located at Scottish Borders (Scotland), United-Kingdom, having the following geographic coordinates: latitude 55°52'16.8"and longitude −2°21'16.2". Penmanshiel wind farm consists of fourteen Senvion MM82 wind turbines, each having a rated power of 2050 KW and hub height of 59 m. These turbines have been commercially operating since September 2016. The spatial representation of the turbines' positions within the wind farm is shown in Figure 5, and the geographic coordinates of each wind turbine is provided in Table 1. Two turbines, 'T01' and 'T05', are selected arbitrarily for experimentation purposes, as depicted on the right-side in Figure 5.

Table 1. Wind turbines' geographic coordinates and their IDs.

Title	ID	Latitude	Longitude	Elevation (m)
Turbine 01	T01	55.902502	−2.306389	212.26
Turbine 02	T02	55.900008	−2.301268	200.46
Turbine 04	T04	55.905943	−2.302690	208.91
Turbine 05	T05	55.903294	−2.298367	201.38
Turbine 06	T06	55.900951	−2.293967	199.03
Turbine 07	T07	55.898741	−2.289856	180.24
Turbine 08	T08	55.907915	−2.297314	200.13
Turbine 09	T09	55.904990	−2.291806	187.04
Turbine 10	T10	55.903032	−2.287585	186.88
Turbine 11	T11	55.900852	−2.282371	204.84
Turbine 12	T12	55.908703	−2.290986	219.31
Turbine 13	T13	55.907026	−2.285887	220.00
Turbine 14	T14	55.905050	−2.281650	219.46
Turbine 15	T15	55.902463	−2.277329	228.15

Figure 5. Penmanshiel wind farm with highlighted target turbines 'T01' and 'T05'. (*Source: Google Maps).*

The SCADA data collected systematically at ten-minute intervals is obtained from [76]. This data includes external and internal variables, which are crucial for accurately representing the operational status of the wind turbines. From the dataset, wind power and wind speed data are used for the period January 2021 to June 2021. After applying data pre-processing techniques mentioned in Section 2.1, the dataset is split into a training, validation, and test set with a ratio of 3:1:1 in chronological order.

4. Experimental setup

In this section, we focused on the training configuration and experimental setup. The model was trained using the mean squared error loss function with a learning rate of $\tau = 0.0001$ and a dropout rate of $r = 0.05$. A batch size of 32 was used, and training was conducted for a single epoch, during which multiple optimization iterations were performed over mini-batches. A sliding-window mechanism with a window size of 32 and a step size of 1 was applied to generate temporally overlapping samples, enabling the model to learn effectively from the available data. The encoder received historical wind-speed sequences together with projected wind speeds of selected surrounding turbines, while the decoder was fed historical wind power to generate multi-step predictions. The key architectural parameters of the VARIMA and Transformer components are summarized in Table 2. A Transformer architecture consisting of two encoders and a single decoder was employed. This configuration has been widely adopted across applications such as short-term load forecasting [77], Time Series Forecasting [78], day-ahead photovoltaic power forecasting [79], tropical cyclone track, and intensity predictions [80,81].

The wind power forecasting experiments in this study were conducted on a PC with 12th Gen Intel(R) Core (TM) i9-12900K 3.20 GHz CPU, 32 GB RAM, and NVIDIA GeForce RTX 4070 GPU. The implementation utilized PyTorch deep learning framework version 1.10.0 + cu102 and Python version 3.9.13.

Table 2. Key parameters of the VARIMA and Transformer components.

Component	Parameter	Value/Strategy	Description
VARIMA	Model order	(p, d, q) selected via AIC/BIC	Lag orders chosen on training data
	Forecast horizon	3 steps	Multi-step wind speed prediction
	Rolling mechanism	Sliding window	Model updated using recent observations
	Window length	32	Length of historical wind-speed input
Transformer	Encoder layers	2	Spatial-temporal feature encoding
	Decoder layers	1	Multi-step power forecasting
	Hidden dimension	512	Embedding and attention dimension
	Attention type	Multi-head self-attention	Captures spatial and temporal dependencies
	Activation function	Gaussian Error Linear Unit (GELU)	Nonlinear transformation
	Input sequence length	32	Past wind-speed time steps
	Decoder label length	16	Historical wind power input
	Output horizon	3 steps	Multi-step wind power prediction

4.1. Evaluation metrics

The performance of the prediction models was assessed using the following metrics:

i. Mean squared error (MSE): A measure of deviation of differences between the measured and model predicted values.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2 \tag{12}$$

ii. Mean absolute error (MAE): The measure of absolute summation of total differences between the measured and model predicted values.

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|\hat{y}_i - y_i| \tag{13}$$

iii. Correlation coefficient (R): Measures the relative strength of the linear relationship between the measured and model predicted values.

$$R = \frac{\sum_{i=1}^{M}\left(\hat{y}_i - \overline{\hat{y}_i}\right)(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{M}\left(\hat{y}_i - \overline{\hat{y}_i}\right)^2 (y_i - \overline{y_i})^2}} \quad , \tag{14}$$

where y_i denotes the true value, $\hat{y}_i$ denotes the predicted value, and M denotes the number of data samples. It should also be noted that R is unaffected by the normalization of wind speed and consistently remains the same. MSE and MAE are frequently employed metrics to quantify the

differences between predicted and actual values, whereas R-metric evaluates the model goodness of fit, indicating how well the model predicts the observed data.

5. Results and discussion

In this section, we cover the results and discussion considering different scenarios. This is initiated first by establishing the temporal Transformer (TT) model. The significance of effective selection of spatial information is established through two separate experiments, i.e., spatio-temporal Transformer (STT) and DTW based spatio-temporal Transformer (DTW-STT). Once the latter approach outperforms the previous one, we update wind speed sequence with VARIMA predicted wind speeds to conform the superiority of DVTransformer. We conclude this section with a thorough comparison of DVTransformer with the state-of-the-art Fast Fourier Transformer (FFTransformer), Informer, Spatio-temporal Long Short-Term Memory (ST-LSTM), and Spatio-temporal Multi-Layer Perceptron (ST-MLP). In the following subsections, we provide details of these experiments.

5.1. Temporal Transformer (TT) model

The Temporal based model employs the features (wind speed and wind power) of the same turbine for model development and evaluation. Evaluation metrics for TT-based models from the 1-step to 3-steps forecast is provided in Table 3. The trend of error suggests that the accuracy of the model is improving as we are moving from 3-steps to 1-step prediction, since the proposed framework is predicting multi steps simultaneously with no accumulated error. Moreover, average MSE and MAE for the 1- to 3- steps ahead prediction are 0.093 and 0.222 for turbine 'T01' and 0.096 and 0.219 for turbine 'T05', respectively. These errors will be subsequently compared with STT models and DVTransformer to signify the underlying spatial information dependencies among wind turbines.

Table 3. Error metrics for temporal, spatio-temporal, and DVTransformer methods.

Turbine	**Variations**		***1 step ahead forecast***			***2 steps ahead forecast***			***3 steps ahead forecast***			***Average***	
ID		**Epochs**	**MSE**	**MAE**	**R**	**MSE**	**MAE**	**R**	**MSE**	**MAE**	**R**	**MSE**	**MAE**
T01	TT	1	0.066	0.181	0.951	0.100	0.240	0.908	0.114	0.244	0.902	0.093	0.222
	STT	1	0.177	0.364	0.812	0.245	0.430	0.701	0.262	0.441	0.675	0.228	0.412
	DTW-STT	1	0.061	**0.170**	0.956	0.089	0.214	0.925	0.111	0.238	0.904	0.087	0.207
	DVTransformer	**1**	**0.060**	**0.170**	**0.957**	**0.087**	**0.210**	**0.929**	**0.109**	**0.234**	**0.913**	**0.085**	**0.205**
T05	TT	1	0.068	0.183	0.936	0.102	0.235	0.888	0.118	0.240	0.882	0.096	0.219
	STT	1	0.271	0.465	0.627	0.346	0.528	0.454	0.369	0.535	0.399	0.329	0.509
	DTW-STT	1	0.070	0.191	0.945	0.094	0.213	0.913	0.116	0.238	0.892	0.093	0.214
	DVTransformer	**1**	**0.066**	**0.181**	**0.950**	**0.091**	**0.210**	**0.920**	**0.114**	**0.235**	**0.902**	**0.090**	**0.208**

5.2. Spatio-temporal Transformer (STT) model

The temporal-based Transformer model, which is trained using the wind speed and power data of the target turbine, demonstrates limited accuracy in forecasting wind power. To address this limitation, spatio-temporal-based Transformer models are developed. These spatio-temporal models incorporate

features extracted from surrounding turbines to enhance the prediction accuracy of wind power for the target turbine.

Two spatio-temporal modeling approaches are evaluated in this study. In the first approach (STT), historical wind speed data from all surrounding turbines apart from the target turbine, along with the wind power data of the target turbine, are utilized to predict the future wind power output of the target turbine.

In the second approach (DTW-STT), we employ historical wind speed data from the selected turbines, chosen based on ascending DTW distance metrics, in conjunction with the target turbine wind power data for prediction purposes. Initially, to identify the number of neighboring turbines empirically, a sensitivity analysis is conducted by varying the number of neighboring turbines 'k'. The value of 'k' varies from 2 to 13, and the corresponding forecasting performance is evaluated. The results are presented in Table 4 for turbines T01 and T05. It is observed that selecting three surrounding turbines (k = 3) yields the lowest error values across MSE and MAE metrics. This indicates that including a moderate number of nearby turbines provides sufficient contextual information while avoiding overfitting or noise from less correlated turbines.

Table 4. Error metrics by varying the number of surrounding wind turbines (k).

Turbine ID	Error	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10	k = 11	k = 12	k = 13
Turbine	MSE	0.090	**0.085**	0.110	0.100	0.090	0.110	0.100	0.090	0.090	0.130	0.092	0.184
T01	MAE	0.220	**0.205**	0.250	0.250	0.220	0.260	0.230	0.210	0.224	0.290	0.211	0.265
Turbine	MSE	0.100	**0.090**	0.100	0.130	0.090	0.110	0.110	0.110	0.098	0.110	0.095	0.323
T05	MAE	0.240	**0.208**	0.230	0.280	0.220	0.240	0.250	0.230	0.229	0.240	0.217	0.506

Figure 6 presents a heatmap of the DTW distance metrics computed from the turbines' wind speed data, revealing varying degrees of spatial correlation among the turbines. Based on the DTW scores, target turbine 'T01' has neighbors 'T02', 'T04', and 'T05', while for target turbine 'T05', the selected neighbor turbines are 'T01', 'T04', and 'T06'.

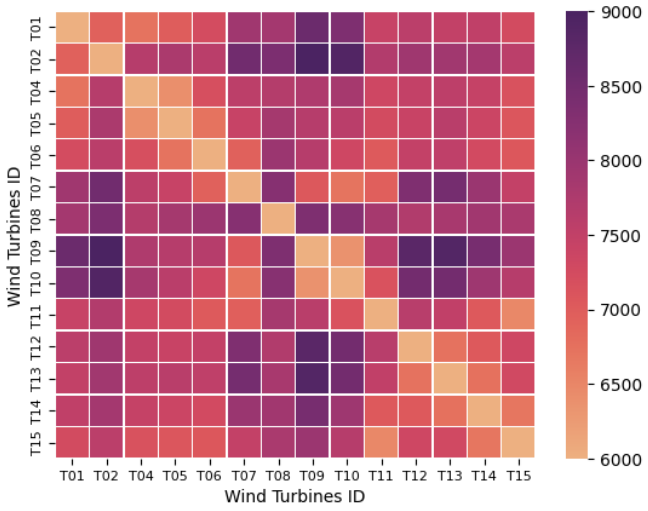

Figure 6. DTW scores heatmap between the wind turbines using their wind speeds.

5.2.1. STT

As given in Table 3, the STT based model shows a significant increase in error metrics compared to the temporal based Transformer model. This indicates that superfluous spatial information is provided to predict the wind power of a target wind turbine and it might not be relevant. This may be overcome by providing only the most relevant spatial information. This is achieved by shortlisting the surrounding wind turbines, which becomes the second scenario of the spatio-temporal based model.

5.2.2. DTW-STT

In the second scenario of the spatio-temporal based Transformer model with selected neighbor turbines, the average MSE and MAE errors are improved to 61.84% and 49.75% for turbine 'T01', respectively, and 71.73% and 57.95% for 'T05', respectively, compared to the STT case. The error trends demonstrate a significant reduction compared to the STT model, enhancing the accuracy of the spatio-temporal model. This also signifies the importance of an appropriate spatio-temporal based model. Compared with the TT model, the MSE and MAE improve by 6.45% and 6.75% for turbine 'T01', respectively, and by 3.12% and 2.28% for turbine 'T05', respectively. This indicates that there is significant information hidden in the surrounding turbines' wind speed time series that cannot be extracted completely when considering only the temporal model. The comprehensive comparison reveals that the particular surrounding atmospheric conditions are critical to the wind power to a greater extent. This further confirms that wind blows in space and passes with the subsequent wind turbines in the wind farm. However, not all the information from the surrounding wind turbines are useful in predicting the wind power, and excessive irrelevant information may degrade the model performance.

5.3. DVTransformer (proposed spatio-temporal method)

To investigate the performance of the proposed method, wind turbines 'T01' and 'T05' are selected for uniform comparison. The proposed model is developed, trained, and evaluated on the same dataset. The results indicate that DVTransformer exhibits superior forecasting accuracy compared to the temporal and spatio-temporal models.

Table 3 presents the evaluation metrics of each forecasting model across 1-step to 3-step forecasting horizons. The error metrics indicate that there is significant improvement in wind power forecasting for the wind turbines using the proposed methodology compared to the TT, STT, and DTW-STT models. For turbine 'T01', MSE and MAE are improved by 2.29% and 0.96%, respectively, compared to DTW-STT. Similarly, MSE and MAE are improved to 3.22% and 2.80%, respectively, for turbine 'T05', which reveals a good agreement between the forecasted and actual wind power. Even though the proposed model reports similar performance in terms of MAE for the single-step forecast compared to DTW-STT for turbine 'T01', it shows significant improvement in MSE. Due to the heavier penalty on larger errors, the MSE indicates that the DTW-STT method has fewer minor errors on average but a higher incidence of substantial mispredictions compared to the proposed method for single-step forecasts. However, for turbine 'T05', all metrics demonstrate effective performance from the DVTransformer.

The improvement in accuracy can be attributed to three major factors. First, the parallelized input structure enhances the efficiency of model training while improving the capacity to internally capture

more distant correlations between sequences. Second, the implementation of multi-head attention in the Transformer effectively captures the correlations between the time series of surrounding wind turbines, which is crucial for accurate predictions. Finally, the key reason for the superior performance of the DVTransformer is the augmentation of predictive strategy through the inclusion of forecasted wind speed time series.

The actual and forecasted time series for turbine 'T05' of the temporal and DVTransformer for the complete test dataset is shown in Figure 7(a), whereas Figure 7(b) shows arbitrary 200 timestamps for better visibility and comparison. Since the proposed model performs multi-step ahead forecasting, 2-steps and 3-steps predicted values are compared with actual time series and are shown in Figure 7(c) and (d), respectively. From these figures, it is evident that in the peak and trough regions of the wind power curve, there is a noticeable deviation between the predicted values and the actual values for the temporal based Transformer model. In contrast, DVTransformer comprehensively captures dependencies among wind turbines, resulting in forecasted values that closely align with the actual output curve. This is also reflected from improved R scores, demonstrating superior forecasting accuracy.

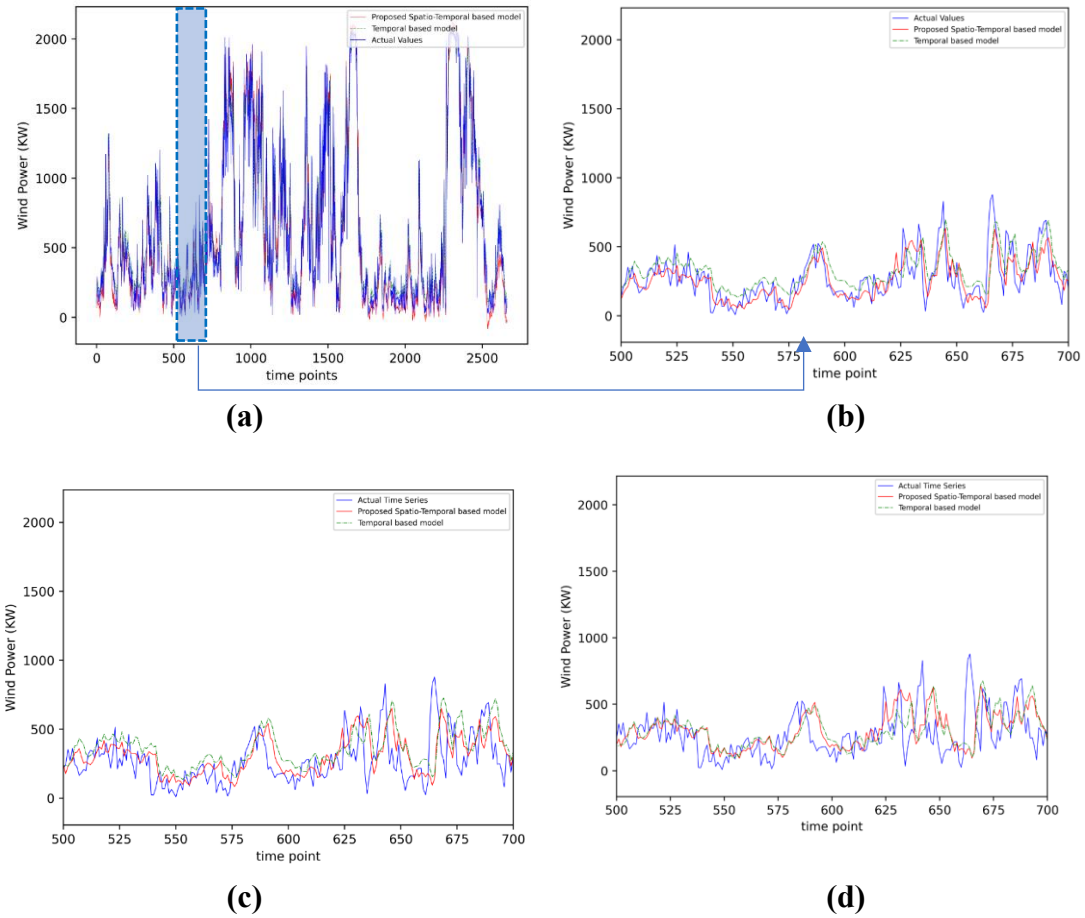

Figure 7. Performance of models for Turbine T05 (a) on the test data; (b) 1-step ahead prediction; (c) 2-steps ahead prediction; and (d) 3-steps ahead prediction.

Following the outlined procedure for selecting the surrounding wind turbines and providing forecasted wind speeds of these turbines at the encoder to forecast the wind power of the target turbine, the predicted future wind powers achieve higher accuracy compared to the original values. In general, incorporating these predicted future wind speeds into historical data has notably enhanced the forecasting performance of the model.

5.4. Comparative analysis with other baselines

To validate the superiority of DVTransformer, the comparative analysis is conducted with baselines and state-of-the-art methods of spatio-temporal forecasting, which includes the FFTransformer [82], Informer [83], ST-LSTM [82], and ST-MLP [82]. For uniform comparison, the same wind turbines are selected to examine the performance of the proposed method. Moreover, comparative models are trained and evaluated on the same dataset with the identical data splitting ratio as the proposed method. Regarding hyperparameters, the decomposition level is set to 4 for FFTransformer. Moreover, ST-LSTM is used in the encoder-decoder setting for multi-step forecasting with 2 encoders and 1 decoder. The comparative models are trained under the 1-epoch and 10-epochs training regime, while the DVTransformer is restricted to 1-epoch of training. For completely trained (10-epochs training) models, an early stopping criterion is utilized up to a patience level of 3.

Table 5 illustrates the experimental results for the proposed method and other state-of-the-art baseline models with the best results highlighted in bold. For 3-step-ahead forecasting, the DVTransformer attains an average MSE of 0.085 for turbine T01, yielding performance improvements of 34.92%, 24.77%, 2.20%, and 27.34% over FFTransformer, Informer, ST-LSTM, and ST-MLP, respectively. Likewise, for turbine T05, the proposed approach records an average MSE of 0.090, corresponding to MSE reductions of 38.45%, 23.08%, 2.93%, and 25.96% for FFTransformer, Informer, ST-LSTM, and ST-MLP, respectively. It is evident from these results that DVTransformer consistently achieves the most accurate results across all forecasting horizons, demonstrating its competitive performance in multi-step wind power forecasting. This success can be attributed by the incorporation of spatial information through a DTW-based metric, which enables the integration of valuable data from surrounding wind turbines, ensuring a reliable understanding of wind conditions. Additionally, the multi-head attention mechanism enables Transformer neural networks to effectively learn short- and long-term dependencies within temporal sequences. Moreover, the inclusion of future knowledge of surrounding wind conditions further enhances the model performance. In general, the effective spatial and temporal modeling provided by the proposed method leads to superior wind power forecasting compared to traditional approaches.

From Table 5, it can be observed that under 1-epoch of training, DVTransformer significantly outperformed the FFTransformer model for all forecasting horizon. For target turbine T01, the proposed model reduces the MSE by 26.22%, 35.82%, and 38.22% and MAE by 17.97%, 20.97%, and 23.02% in the 1-, 2-, and 3-step ahead predictions, respectively. The same pattern is observed for turbine T05, with a reduction in MSE by 53.51%, 32.21%, and 30.59% and a reduction in MAE by 43.28%, 22.47%, and 23.71% in the 1-, 2-, and 3-step ahead prediction, respectively. Considering the computational time, the FFTransformer is significantly slower than other models since it relies on computing FFTs, which are fairly slow and take longer to compute forecasts.

DVTransformer achieves noteworthy improvements and outperforms the Informer model across all forecasting horizons under the 1-epoch training regime, as listed in Table 5. For turbine T01, it

reduces the MSE by 19.02%, 30.90%, and 22.31%, and the MAE by 21.80%, 28.12%, and 23.50% in the 1-, 2-, and 3-step ahead predictions, respectively. A similar trend is observed for turbine T05, with MSE reductions of 19.83%, 30.89%, and 17.60%, and MAE reductions of 22.16%, 26.53%, and 19.23% for the 1-, 2-, and 3-step ahead predictions, respectively. In general, DVTransformer attains better results under the 1-epoch training regime than the other variants of Transformer-based models; i.e., FFTransformer and Informer models for the multi-step forecasts.

Table 5. Comparison of error metrics with other baselines.

			1 step ahead forecast			*2 steps ahead forecast*			*3 steps ahead forecast*			*Average*	
Turbine ID	**Method**	**Epochs**	**MSE**	**MAE**	**R**	**MSE**	**MAE**	**R**	**MSE**	**MAE**	**R**	**MSE**	**MAE**
T01	**DVTransformer**	1	0.060	0.170	**0.957**	**0.087**	**0.210**	**0.929**	**0.109**	**0.234**	**0.913**	**0.085**	**0.205**
	FFTransformer	1	0.081	0.207	0.940	0.135	0.266	0.856	0.177	0.304	0.790	0.131	0.259
	Informer	1	0.074	0.218	0.930	0.126	0.292	0.843	0.141	0.306	0.832	0.113	0.272
	ST-LSTM	1	0.062	0.184	0.944	0.089	0.215	0.919	0.111	0.251	0.887	0.087	0.217
	ST-MLP	1	0.090	0.220	0.933	0.116	0.247	0.906	0.146	0.297	0.856	0.117	0.255
	FFTransformer	10	**0.056**	0.165	0.955	0.091	0.213	0.923	0.112	0.240	0.902	0.086	0.206
	Informer	10	0.056	**0.164**	0.956	0.092	0.219	0.917	0.112	0.245	0.897	0.087	0.209
	ST-LSTM	10	0.062	0.180	0.948	0.090	0.213	0.923	0.111	0.247	0.896	0.088	0.213
	ST-MLP	10	0.076	0.201	0.942	0.104	0.231	0.916	0.125	0.256	0.892	0.102	0.230
T05	**DVTransformer**	1	0.066	0.181	**0.950**	**0.091**	**0.210**	**0.920**	**0.114**	**0.235**	**0.902**	**0.090**	**0.208**
	FFTransformer	1	0.141	0.319	0.854	0.134	0.270	0.891	0.165	0.307	0.854	0.147	0.299
	Informer	1	0.082	0.232	0.909	0.131	0.285	0.850	0.139	0.290	0.838	0.117	0.269
	ST-LSTM	1	0.069	0.191	0.934	0.095	0.220	0.908	0.115	0.254	0.874	0.093	0.222
	ST-MLP	1	0.090	0.211	0.931	0.115	0.246	0.897	0.161	0.305	0.845	0.122	0.254
	FFTransformer	10	0.060	0.167	0.947	0.095	0.217	0.910	0.115	0.241	0.885	**0.090**	**0.208**
	Informer	10	**0.057**	**0.163**	**0.950**	0.102	0.225	0.907	0.119	0.245	0.888	0.093	0.211
	ST-LSTM	10	0.069	0.182	0.941	0.097	0.216	0.913	0.116	0.243	0.888	0.094	0.213
	ST-MLP	10	0.087	0.204	0.933	0.108	0.230	0.910	0.128	0.254	0.885	0.108	0.230

For wind power time series forecasting, Long Short-Term Memory (LSTM) networks are generally considered an effective method. Consequently, the DVTransformer is also experimentally compared with the ST-LSTM model. Compared to the ST-LSTM model, the DVTransformer shows marginal improvements for the multi-step forecasts under the 1-epoch training regime. For turbine T01, it reduces MSE by 4.04%, 2.22%, and 1.15%, and MAE by 7.54%, 2.44%, and 6.89% in the 1-, 2-, and 3-step ahead forecasts, respectively. Turbine T05 also follows this pattern, with MSE reductions of 4.89%, 3.88%, and 0.99%, and MAE reductions of 5.43%, 4.68%, and 7.50% in the 1-, 2-, and 3-step ahead forecasts, respectively, confirming the superiority of the DVTransformer over ST-LSTM. Additionally, the ST-LSTM model shows superior performance than the ST-MLP model for the 1-, 2-, and 3-steps forecasts. This indicates that the ST-LSTM architecture is better at encoding the useful information, resulting in the improved performance of the ST-LSTM model compared to the ST-MLP.

DVTransformer also demonstrates superior performance over the ST-MLP model across all forecasting horizons. For turbine T01, MSE is lowered by 33.53%, 25.20%, and 25.22%, while MAE decreases by 22.62%, 14.82%, and 21.38% for 1-, 2-, and 3-step ahead predictions, respectively.

Turbine T05 exhibits a similar trend, with MSE reductions of 26.83%, 21.22%, and 28.87%, and MAE reductions of 14.31%, 14.85%, and 23.21% for 1-, 2-, and 3-step ahead predictions, respectively.

As shown in Table 5, the DVTransformer trained for only a single epoch achieves an accuracy comparable to that of the fully trained (10-epoch) baseline models, thereby demonstrating its superior learning efficiency. Figures 8 and 9 illustrate representative predictions for 1-step and 3-step forecasting horizons at arbitrarily selected timestamps for turbine T05. In the multi-step forecasting setting, the proposed model generates more diverse and accurate temporal patterns that closely follow the actual wind power trajectory across all time steps, compared to the competing models. Furthermore, the results indicate that the performance of the competing models degrades as the forecasting horizon increases (see Figure 9). In contrast, the DVTransformer consistently outperforms the other models across all prediction intervals. This behavior highlights the robustness of the proposed approach and its strong generalization capability, particularly in longer-term forecasting scenarios.

To evaluate the physical implications of forecasting results in the context of wind energy production, an inverse transformation is applied to actual and predicted wind power outputs, followed by the calculation of MAEs. The results yield rough estimates of average power errors across models. As shown in Table 6, these findings are consistent with those discussed in Table 5, providing enhanced interpretability regarding the effects of differing predictive performances and the associated risks linked to alternative models.

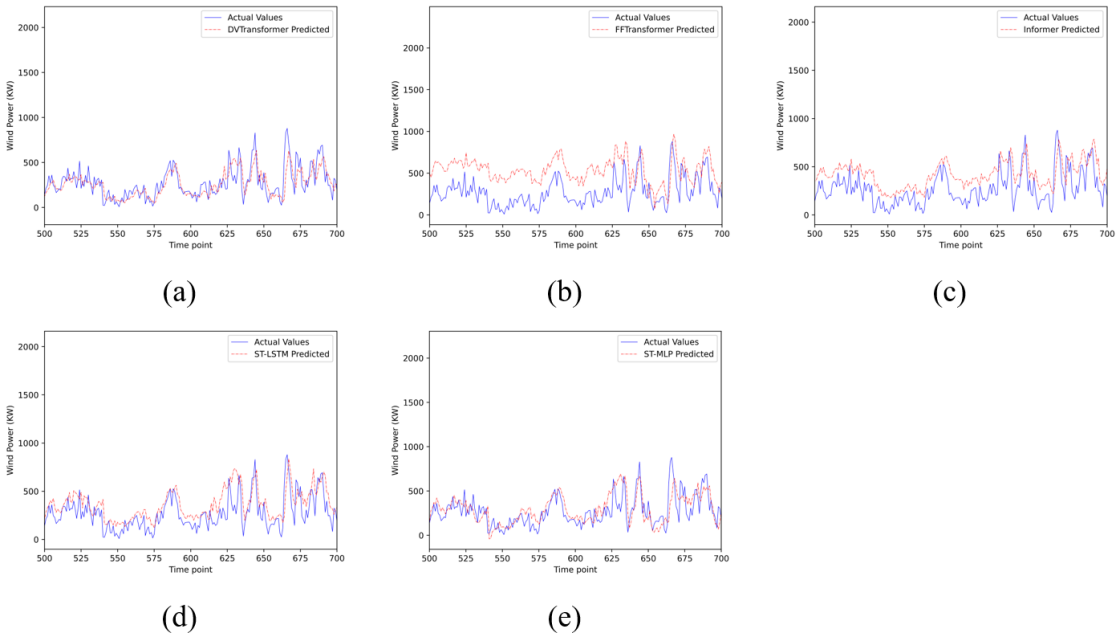

Figure 8. One step ahead prediction result for turbine T05 using 1-epoch training: (a) DVTransformer, (b) FFTransformer, (c) Informer, (d) ST-LSTM, and (e) ST-MLP.

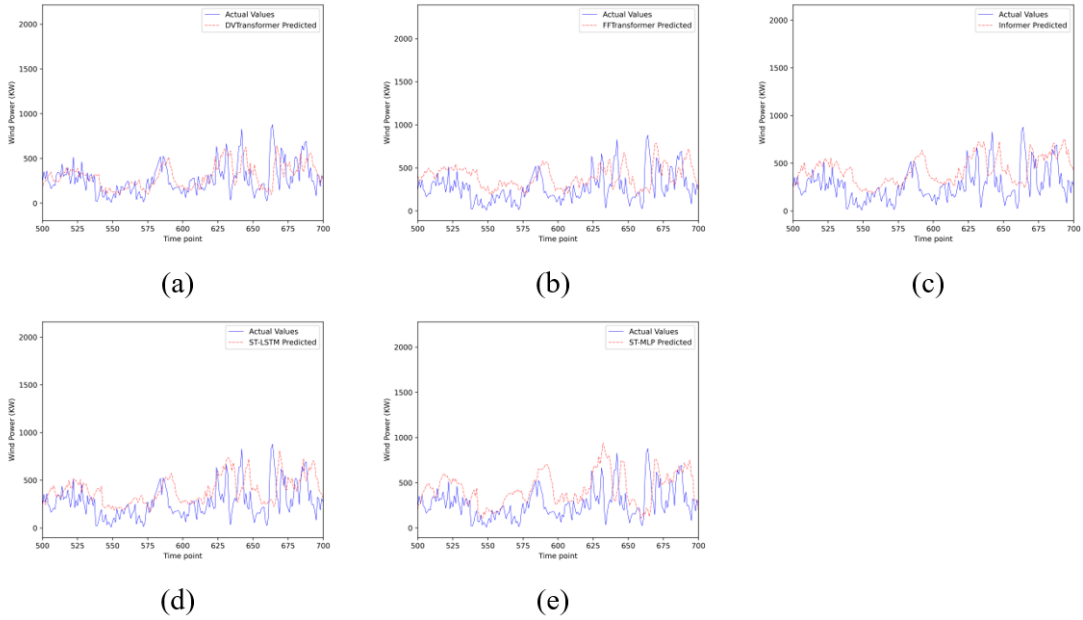

Figure 9. Three-steps ahead prediction results for turbine T05 using 1-epoch training: (a) DVTransformer, (b) FFTransformer, (c) Informer, (d) ST-LSTM, and (e) ST-MLP.

Table 6. Performance evaluation in KW and percentage improvement.

			1-step ahead (10 mins)		2-steps ahead (20 mins)		3-steps ahead (30 mins)	
Turbine ID	**Method**	**Epochs**	**MAE [kW]**	**% Improve**	**MAE [kW]**	**% Improve**	**MAE [kW]**	**% Improve**
T01	**DVTransformer**	1	122.24	--	150.89	--	167.92	--
	FFTransformer	1	149.02	**17.97**	190.95	**20.98**	218.15	**23.03**
	Informer	1	156.33	**21.81**	209.94	**28.13**	219.56	**23.52**
	ST-LSTM	1	132.21	**7.54**	154.67	**2.45**	180.35	**6.89**
	ST-MLP	1	157.98	**22.63**	177.16	**14.83**	213.59	**21.38**
	FFTransformer	10	118.52	**-3.13**	153.28	**1.56**	172.11	**2.44**
	Informer	10	117.99	**-3.60**	157.52	**4.21**	175.70	**4.43**
	ST-LSTM	10	129.44	**5.56**	153.16	**1.48**	177.20	**5.24**
	ST-MLP	10	144.56	**15.44**	166.30	**9.27**	183.78	**8.63**
T05	**DVTransformer**	1	128.87	--	149.34	--	167.06	--
	FFTransformer	1	227.23	**43.29**	192.64	**22.47**	219.01	**23.72**
	Informer	1	165.57	**22.17**	203.28	**26.53**	206.85	**19.24**
	ST-LSTM	1	136.28	**5.44**	156.68	**4.69**	180.62	**7.51**
	ST-MLP	1	150.40	**14.31**	175.40	**14.85**	217.57	**23.22**
	FFTransformer	10	118.90	**-8.38**	154.36	**3.25**	171.74	**2.72**
	Informer	10	116.26	**-10.85**	160.20	**6.78**	174.73	**4.39**
	ST-LSTM	10	129.52	**0.50**	153.71	**2.84**	172.79	**3.31**
	ST-MLP	10	145.33	**11.32**	164.09	**8.99**	181.19	**7.80**

The DVTransformer consistently demonstrates superior stability in forecasting across all time horizons compared to the baseline models. This empirical evidence highlights the improved effectiveness of the multi-head attention mechanism integrated into the Transformer architecture, particularly in our problem. The proposed enhancement of projected wind conditions significantly increases the model ability to capture and process temporal and spatial data across points in time and space.

To ensure fairness in evaluating convergence and performance, a training-budget sensitivity analysis comparing DVTransformer trained for a single epoch with a completely trained Transformer is performed. As shown in Table 7, the DVTransformer achieves prediction accuracy comparable to the trained Transformer across all turbines while reducing the computational time by 6-7 times. This indicates that the proposed model rapidly converges to a sufficiently optimal parameter region within one epoch.

These findings suggest that additional training epochs provide limited accuracy gains relative to their computational cost, thereby validating the single-epoch protocol.

Table 7. Training-budget sensitivity analysis.

Model	Evaluation metrics	T01	T02	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T15
Transformer	MSE	0.085	0.083	0.083	0.087	0.086	0.082	0.080	0.084	0.080	0.079	0.089	0.088	0.080	0.081
(Completely trained)	MAE	0.203	0.202	0.203	0.206	0.203	0.195	0.202	0.198	0.194	0.196	0.212	0.211	0.201	0.199
	Training time (s)	280.4	197.2	174.4	236.4	248.1	199.4	180.0	205.4	230.4	248.7	195.1	234.7	180.7	224.5
DVTransformer	MSE	0.085	0.083	0.087	0.090	0.090	0.090	0.086	0.089	0.095	0.082	0.104	0.102	0.090	0.095
(1-epoch trained)	MAE	0.205	0.203	0.210	0.208	0.209	0.208	0.211	0.208	0.228	0.200	0.231	0.232	0.216	0.221
	Training time (sec)	38.78	38.21	38.34	39.46	38.00	37.73	37.45	37.46	39.89	37.62	38.87	38.46	37.86	38.79

5.5. Ablation study

To systematically evaluate the factors contributing to the predictive performance of our framework, an ablation study is conducted to evaluate the contribution of VARIMA predicted wind speeds. Table 8 summarizes the results for seven selected turbines (T02, T04, T09, T11, T12, T13, and T14). Across these turbines, the inclusion of VARIMA consistently reduces both error metrics. As shown in Figure 10, the consistency across turbines highlights that even short-horizon VARIMA forecasts provide informative temporal indications that enhance the transformer's ability to capture turbine-level dynamics.

Table 8. Error metrics for the Transformer (without VARIMA) and DVTransformer.

		Penmanshiel Wind Turbine IDs (with k = 3)						
Model	**Error**	**T02**	**T04**	**T09**	**T11**	**T12**	**T13**	**T14**
DTW + Transformer (Without VARIMA)	MSE	0.086	0.089	0.096	0.084	0.105	0.114	0.098
	MAE	0.207	0.214	0.218	0.203	0.234	0.254	0.232
DVTransformer (Including VARIMA)	MSE	0.083	0.087	0.089	0.082	0.104	0.102	0.090
	MAE	0.203	0.210	0.208	0.200	0.231	0.232	0.216

This ablation study validates the tangible benefits of integrating short-term statistical forecasts into a deep learning-based wind power prediction model.

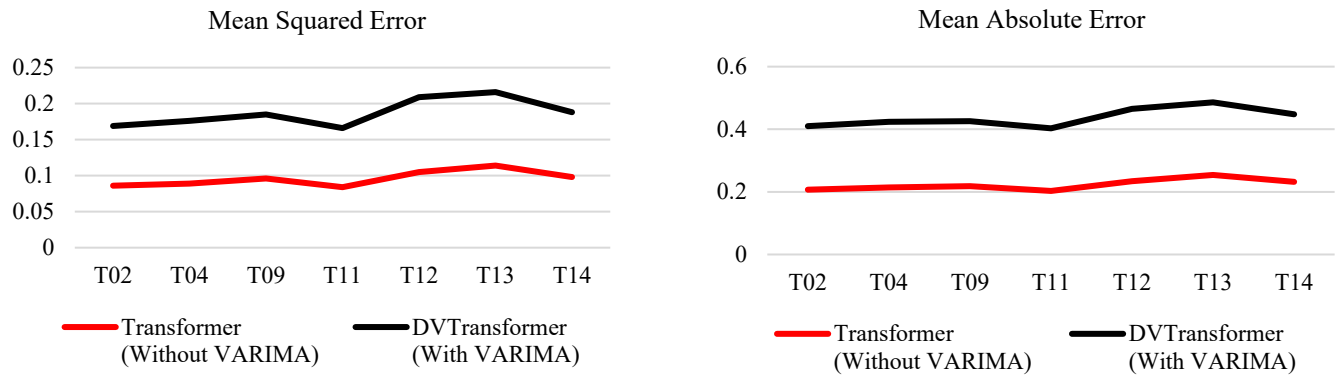

Figure 10. MAE and MSE scores by removing and considering VARIMA predictions.

6. Conclusions

In recent years, Transformer-based models have dominated sequence-based deep learning, but their application in wind forecasting remains limited. In this study, we introduce a novel DVTransformer framework for multi-step wind power forecasting, demonstrating superior performance compared to baseline models. The method incorporates a DTW-based evaluation scheme to optimize the selection of influential turbines for spatial information, rather than using data from all surrounding turbines. The model effectively integrates historical and predicted wind conditions using VARIMA, achieving highly accurate predictions in experiments with real-world wind farm data. Additionally, the model flexibility enables easy adaptation to forecasting applications.

7. Limitations and future work

Despite the promising results, this study has certain limitations that provide directions for future research. The proposed framework employs a fixed set of three surrounding wind turbines to model spatial dependencies; however, the optimal number of contributing turbines may vary across wind farms depending on factors such as farm layout, terrain complexity, wind direction, upstream-downstream interactions, and prevailing meteorological conditions. Therefore, in future work, we will

focus on developing dynamic turbine selection strategies that adaptively determine the number and identity of relevant surrounding turbines using spatial proximity, correlation strength, or data-driven relevance measures. In addition, the Transformer layer configuration and hyperparameters could be further optimized through systematic approaches, such as greedy line search, as performed in [35]. Finally, the integration of additional exogenous variables, including atmospheric pressure, temperature, and turbulence intensity, will further enhance forecasting accuracy and robustness under diverse operating conditions.

Use of AI tools declaration

During the preparation of this work the authors used ChatGPT (OpenAI) in order to improve the clarity, grammar, and overall readability of the manuscript. 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.

Data availability

The SCADA data used in this study are publicly available and can be accessed from URL: https://zenodo.org/records/5946808.

Acknowledgments

The authors would like to acknowledge Penmanshiel wind farm for their public reports and open access SCADA data.

This study was supported by the NED University of Engineering & Technology, Karachi, Pakistan, and funded by Ministry of Science & Technology (MoST) Endowment Fund, Government of Pakistan (Grant No. Acad/50(48)/87004). The authors also acknowledge the use of Google Maps for visualization of the study area.

Author contributions

Syed Muhammad Rashid Hussain: Conceptualization, Methodology, Software, Validation, Formal, analysis, Investigation, Writing—original draft; Mirza Muhammad Ali Baig: Resources, Supervision, Writing—review & editing; Muhammad Uzair Yousuf: Resources, Supervision, Writing—review & editing.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wang M, Yao M, Wang S, et al. (2021) Study of the emissions and spatial distributions of various power-generation technologies in China. *J Environ Manage* 278: 111401. https://doi.org/10.1016/j.jenvman.2020.111401
2. Yousuf MU, Malik MH, Umair M (2024) 4E analysis of solar photovoltaic, wind, and hybrid power systems in southern Pakistan: Energy, exergy, economic, and environmental perspectives. *Sci Technol Energy Trans* 79: 94. https://doi.org/10.2516/stet/2024088
3. IRENA (2024) Renewable capacity statistics 2024, International Renewable Energy Agency, Abu Dhabi. Available from: https://www.irena.org/Publications/2024/Mar/Renewable-capacity-statistics-2024.
4. Medina C, González G (2022) Transmission grids to foster high penetration of large-scale variable renewable energy sources—A review of challenges, problems, and solutions. *Int J Renewable Energy Res* 12: 146–169. https://doi.org/10.20508/ijrer.v12i1.12738.g8400
5. Asiaban S, Kayedpour N, Samani AE, et al. (2021) Wind and solar intermittency and the associated integration challenges: A comprehensive review including the status in the Belgian power system. *Energies* 14: 2630. https://doi.org/10.3390/en14092630
6. Martinot E (2016) Grid integration of renewable energy: Flexibility, innovation, and experience. *Annu Rev Environ Resour* 41: 223–251. https://doi.org/10.1146/annurev-environ-110615-085725
7. Yan J, Möhrlen C, Göçmen T, et al. (2022) Uncovering wind power forecasting uncertainty sources and their propagation through the whole modelling chain. *Renewable Sustainable Energy Rev* 165: 112519. https://doi.org/10.1016/j.rser.2022.112519
8. Wang Y, Zou R, Liu F, et al. (2021) A review of wind speed and wind power forecasting with deep neural networks. *Appl Energy* 304: 117766. https://doi.org/10.1016/j.apenergy.2021.117766
9. Santhosh M, Venkaiah C, Vinod Kumar D (2020) Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review. *Eng Rep* 2: e12178. https://doi.org/10.1002/eng2.12178
10. Liu H, Chen C, Lv X, et al. (2019) Deterministic wind energy forecasting: A review of intelligent predictors and auxiliary methods. *Energy Convers Manage* 195: 328–345. https://doi.org/10.1016/j.enconman.2019.05.020
11. Yousuf MU, Al-Bahadly I, Avci E (2022) Statistical wind speed forecasting models for small sample datasets: Problems, improvements, and prospects. *Energy Convers Manage* 261: 115658. https://doi.org/10.1016/j.enconman.2022.115658
12. Yildiz C, Acikgoz H, Korkmaz D, et al. (2021) An improved residual-based convolutional neural network for very short-term wind power forecasting. *Energy Convers Manage* 228: 113731. https://doi.org/10.1016/j.enconman.2020.113731
13. Hossain MA, Gray E, Lu J, et al. (2023) Optimized forecasting model to improve the accuracy of very short-term wind power prediction. *IEEE Trans Ind Inf* 19: 10145–10159. https://doi.org/10.1109/TII.2022.3230726
14. Zhang Y, Li Y, Zhang G (2020) Short-term wind power forecasting approach based on Seq2Seq model using NWP data. *Energy* 213: 118371. https://doi.org/10.1016/j.energy.2020.118371
15. Xiong B, Lou L, Meng X, et al. (2022) Short-term wind power forecasting based on Attention Mechanism and Deep Learning. *Electr Power Syst Res* 206: 107776. https://doi.org/10.1016/j.epsr.2022.107776

16. Shirzadi N, Nasiri F, El-Bayeh C, et al. (2022) Optimal dispatching of renewable energy-based urban microgrids using a deep learning approach for electrical load and wind power forecasting. *Int J Energy Res* 46: 3173–3188. https://doi.org/10.1002/er.7374
17. Chen C, Liu H (2020) Medium-term wind power forecasting based on multi-resolution multi-learner ensemble and adaptive model selection. *Energy Convers Manage* 206: 112492. https://doi.org/10.1016/j.enconman.2020.112492
18. Ahmadi A, Nabipour M, Mohammadi-Ivatloo B, et al. (2020) Long-term wind power forecasting using tree-based learning algorithms. *IEEE Access* 8: 151511–151522. https://doi.org/10.1109/ACCESS.2020.3017442
19. Suárez-Cetrulo AL, Burnham-King L, Haughton D, et al. (2022) Wind power forecasting using ensemble learning for day-ahead energy trading. *Renewable Energy* 191: 685–698. https://doi.org/10.1016/j.renene.2022.04.032
20. Guo Z, Zhao F, Wang B, et al. (2024) Medium to long term wind power prediction based on improved mRMR algorithm and GWO-LSTM algorithm. *2024 6th International Conference on Energy Systems and Electrical Power (ICESEP)*, 137–141. https://doi.org/10.1109/ICESEP62218.2024.10651978
21. Yousuf MU, Al-Bahadly I, Avci E (2021) Short-term wind speed forecasting based on hybrid MODWT-ARIMA-Markov model. *IEEE Access* 9: 79695–79711. https://doi.org/10.1109/ACCESS.2021.3084536
22. Yousuf MU, Al-Bahadly I, Avci E (2019) Current perspective on the accuracy of deterministic wind speed and power forecasting. *IEEE Access* 7: 159547–159564. https://doi.org/10.1109/ACCESS.2019.2951153
23. Jónsson T, Pinson P, Nielsen HA, et al. (2014) Exponential smoothing approaches for prediction in real-time electricity markets. *Energies* 7: 3710–3732. https://doi.org/10.3390/en7063710
24. Yousuf MU, Al-Bahadly I, Avci E (2022) Wind speed prediction for small sample dataset using hybrid first-order accumulated generating operation-based double exponential smoothing model. *Energy Sci Eng* 10: 726–739. https://doi.org/10.1002/ese3.1047
25. Yousuf MU, Al-Bahadly I, Avci E (2021) A modified GM (1, 1) model to accurately predict wind speed. *Sustainable Energy Technol Assess* 43: 100905. https://doi.org/10.1016/j.seta.2020.100905
26. Bhaskar K, Singh SN (2012) AWNN-assisted wind power forecasting using feed-forward neural network. *IEEE Trans Sustainable Energy* 3: 306–315. https://doi.org/10.1109/TSTE.2011.2182215
27. Cassola F, Burlando M (2012) Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output. *Appl Energy* 99: 154–166. https://doi.org/10.1016/j.apenergy.2012.03.054
28. Lahouar A, Slama JBH (2017) Hour-ahead wind power forecast based on random forests. *Renewable Energy* 109: 529–541. https://doi.org/10.1016/j.renene.2017.03.064
29. Li LL, Zhao X, Tseng ML, et al. (2020) Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. *J Cleaner Prod* 242: 118447. https://doi.org/10.1016/j.jclepro.2019.118447
30. Severiano CA, de Lima e Silva PC, Cohen MW, et al. (2021) Evolving fuzzy time series for spatio-temporal forecasting in renewable energy systems. *Renewable Energy* 171: 764–783. https://doi.org/10.1016/j.renene.2021.02.117

31. Zhao Y, Ye L, Li Z, et al. (2016) A novel bidirectional mechanism based on time series model for wind power forecasting. *Appl Energy* 177: 793–803. https://doi.org/10.1016/j.apenergy.2016.03.096
32. Song L, Xie Q, He Y, et al. (2020) Ultra-short-term wind power combination forecasting model based on MEEMD-SAE-Elman. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE*, 1844–1850. https://doi.org/10.1109/ITNEC48623.2020.9084768
33. Li J, Geng D, Zhang P, et al. (2019) Ultra-short term wind power forecasting based on LSTM neural network. *2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), IEEE*, 1815–1818. https://doi.org/10.1109/CIEEC47146.2019.CIEEC-2019625
34. Yu Y, Han X, Yang M, et al. (2019) Probabilistic prediction of regional wind power based on spatiotemporal quantile regression. *IEEE Industry Applications Society Annual Meeting, IEEE*, 1–16. https://doi.org/10.1109/IAS.2019.8911916
35. Sun S, Liu Y, Li Q, et al. (2023) Short-term multi-step wind power forecasting based on spatio-temporal correlations and Transformer neural networks. *Energy Convers Manage* 283: 116916. https://doi.org/10.1016/j.enconman.2023.116916
36. Niu Z, Yu Z, Tang W, et al. (2020) Wind power forecasting using attention-based gated recurrent unit network. *Energy* 196: 117081. https://doi.org/10.1016/j.energy.2020.117081
37. Zhang J, Yan J, Infield D, et al. (2019) Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and Gaussian mixture model. *Appl Energy* 241: 229–244. https://doi.org/10.1016/j.apenergy.2019.03.044
38. Wang Y, Hu Q, Srinivasan D, et al. (2018) Wind power curve modeling and wind power forecasting with inconsistent data. *IEEE Trans Sustainable Energy* 10: 16–25. https://doi.org/10.1109/TSTE.2018.2820198
39. Yan J, Zhang H, Liu Y, et al. (2017) Forecasting the high penetration of wind power on multiple scales using multi-to-multi mapping. *IEEE Trans Power Syst* 33: 3276–3284. https://doi.org/10.1109/TPWRS.2017.2787667
40. Prósper MA, Otero-Casal C, Fernández FC, et al. (2019) Wind power forecasting for a real onshore wind farm on complex terrain using WRF high resolution simulations. *Renewable Energy* 135: 674–686. https://doi.org/10.1016/j.renene.2018.12.047
41. Ezzat AA (2020) Turbine-specific short-term wind speed forecasting considering within-farm wind field dependencies and fluctuations. *Appl Energy* 269: 115034. https://doi.org/10.1016/j.apenergy.2020.115034
42. Kusiak A, Verma A (2012) Monitoring wind farms with performance curves. *IEEE Trans Sustainable Energy* 4: 192–199. https://doi.org/10.1109/TSTE.2012.2212470
43. Lee G, Ding Y, Genton MG, et al. (2015) Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J Am Stat Assoc* 110: 56–67. https://doi.org/10.1080/01621459.2014.977385
44. Gill S, Stephen B, Galloway S (2011) Wind turbine condition assessment through power curve copula modeling. *IEEE Trans Sustainable Energy* 3: 94–101. https://doi.org/10.1109/TSTE.2011.2167164
45. Santos RA (2007) Damage mitigating control for wind turbines: University of Colorado at Boulder. Available from: https://ui.adsabs.harvard.edu/abs/2007PhDT.....91S/abstract.

46. Kragh KA, Hansen MH (2014) Load alleviation of wind turbines by yaw misalignment. *Wind Energy* 17: 971–982. https://doi.org/10.1002/we.1612
47. Howland MF, Lele SK, Dabiri JO (2019) Wind farm power optimization through wake steering. *Proc Natl Acad Sci* 116: 14495–14500. https://doi.org/10.1073/pnas.1903680116
48. Yildirim M, Gebraeel NZ, Sun XA (2017) Integrated predictive analytics and optimization for opportunistic maintenance and operations in wind farms. *IEEE Trans Power Syst* 32: 4319–4328. https://doi.org/10.1109/TPWRS.2017.2666722
49. Byon E, Ntaimo L, Ding Y (2010) Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. *IEEE Trans Reliab* 59: 393–404. https://doi.org/10.1109/TR.2010.2046804
50. Browell J, Gilbert C, McMillan D (2017) Use of turbine-level data for improved wind power forecasting. *IEEE Manchester PowerTech*, *IEEE*, 1–6. https://doi.org/10.1109/PTC.2017.7981134
51. Yakoub G, Mathew S, Leal J (2023) Direct and indirect short-term aggregated turbine-and farm-level wind power forecasts integrating several NWP sources. *Heliyon,* 9. https://doi.org/10.1016/j.heliyon.2023.e21479
52. Deng J, Xiao Z, Zhao Q, et al. (2024) Wind turbine short-term power forecasting method based on hybrid probabilistic neural network. *Energy* 313: 134042. https://doi.org/10.1016/j.energy.2024.134042
53. Su H, Du Y, Che Y, et al. (2025) Hybrid wind power forecasting for turbine clusters: Integrating spatiotemporal WGANs with extreme missing-data resilience. *Sustainability* 17: 9200. https://doi.org/10.3390/su17209200
54. Sopeña JG, Maury C, Pakrashi V, et al. (2022) Turbine-level clustering for improved short-term wind power forecasting. *J Phys: Conf Ser* 2265: 022052. https://doi.org/10.1088/1742-6596/2265/2/022052
55. Hering AS, Genton MG (2010) Powering up with space-time wind forecasting. *J Am Stat Assoc* 105: 92–104. https://doi.org/10.1198/jasa.2009.ap08117
56. Dowell J, Weiss S, Hill D, et al. (2014) Short-term spatio-temporal prediction of wind speed and direction. *Wind Energy* 17: 1945–1955. https://doi.org/10.1002/we.1682
57. Dowell J, Pinson P (2015) Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans Smart Grid* 7: 763–770. https://doi.org/10.1109/TSG.2015.2424078
58. Tastu J, Pinson P, Trombe PJ, et al. (2013) Probabilistic forecasts of wind power generation accounting for geographically dispersed information. *IEEE Trans Smart Grid* 5: 480–489. https://doi.org/10.1109/TSG.2013.2277585
59. Yu C, Yan G, Yu C, et al. (2023) Attention mechanism is useful in spatio-temporal wind speed prediction: Evidence from China. *Appl Soft Comput* 148: 110864. https://doi.org/10.1016/j.asoc.2023.110864
60. Zhen H, Niu D, Yu M, et al. (2020) A hybrid deep learning model and comparison for wind power forecasting considering temporal-spatial feature extraction. *Sustainability* 12: 9490. https://doi.org/10.3390/su12229490
61. Yu G, Liu C, Tang B, et al. (2022) Short term wind power prediction for regional wind farms based on spatial-temporal characteristic distribution. *Renewable Energy* 199: 599–612. https://doi.org/10.1016/j.renene.2022.08.142

62. Das A, Kong W, Leach A, et al. (2023) Long-term forecasting with tide: Time-series dense encoder. *arXiv Preprint arXiv: 230408424*. https://doi.org/10.48550/arXiv.2304.08424
63. Gong Y, Li Z, Zhang J, et al. (2020) Online spatio-temporal crowd flow distribution prediction for complex metro system. *IEEE Trans Knowl Data Eng* 34: 865–880. https://doi.org/10.1109/TKDE.2020.2985952
64. Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Adv Neural Inf Process Syst*. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
65. Karad S, Thakur R (2021) Efficient monitoring and control of wind energy conversion systems using Internet of things (IoT): A comprehensive review. *Environ, Dev Sustainability* 23: 14197–14214. https://doi.org/10.1007/s10668-021-01267-6
66. Frery AC (2023) Interquartile Range. *Encycl Math Geosci*, 664–666. https://doi.org/10.1007/978-3-030-85040-1_165
67. Tascikaraoglu A (2018) Evaluation of spatio-temporal forecasting methods in various smart city applications. *Renewable Sustainable Energy Rev* 82: 424–435. https://doi.org/10.1016/j.rser.2017.09.078
68. Ji T, Wang J, Li M, et al. (2022) Short-term wind power forecast based on chaotic analysis and multivariate phase space reconstruction. *Energy Convers Manage* 254: 115196. https://doi.org/10.1016/j.enconman.2021.115196
69. Woźniak A, Kluczek A, Nycz PD (2024) Approach for identifying the impact of local wind and spatial conditions on wind turbine blade geometry. *Int J Energy Res* 2024: 7310206. https://doi.org/10.1155/2024/7310206
70. Cicilio P, Cotilla-Sanchez E (2019) Evaluating measurement-based dynamic load modeling techniques and metrics. *IEEE Trans Power Syst* 35: 1805–1811. https://doi.org/10.1109/TPWRS.2019.2949722
71. Goh H, Lee S, Chua Q, et al. (2016) Wind energy assessment considering wind speed correlation in Malaysia. *Renewable Sustainable Energy Rev* 54: 1389–1400. https://doi.org/10.1016/j.rser.2015.10.076
72. Shumway RH, Stoffer DS, Shumway RH, et al. (2017) ARIMA models. *Time Ser Anal Appl,* 75–163. https://doi.org/10.1007/978-3-319-52452-8_3
73. Tsay RS (2005) Analysis of financial time series. John Wiley & Sons. https://doi.org/10.1002/0471746193
74. Rusyana A, Rahmati L, Nurhasanah N (2020) Forecasting of revenue, number of plane movements and number of passenger movements at sultan iskandar muda international airport using the VARIMA method. *Proceedings of the 1st International Conference on Statistics and Analytics*. https://doi.org/10.4108/eai.2-8-2019.2290496
75. Ajayi A, Luk PC-K, Lao L, et al. (2023) Energy forecasting model for ground movement operation in green airport. *Energies* 16: 5008. https://doi.org/10.3390/en16135008
76. Plumley C (2022) Penmanshiel wind farm data. *Powered by CERN Data Centre & InvenioRDM*. Availabel from: https://zenodo.org/records/5946808.

77. Liu J, Liu G, Ruan J, et al. (2022) Short-term load forecasting with frequency enhanced decomposed transformer. *2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2), IEEE*, 1766–1771. https://doi.org/10.1109/EI256261.2022.10116459
78. Zhou C, Che C, Wang P, et al. (2024) SCAT: A time series forecasting with spectral central alternating transformers. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 5626–5634. https://doi.org/10.24963/ijcai.2024/622
79. Huang X, Ding X, Han Y, et al. (2025) Day-Ahead photovoltaic power forecasting based on SN-Transformer-BiMixer. *Energies* 18: 4406. https://doi.org/10.3390/en18164406
80. Lin Z, Chu JE, Ham YG (2025) Enhancing tropical cyclone track and intensity predictions with the OWZP-Transformer model. *npj Artif Intell* 1: 33. https://doi.org/10.1038/s44387-025-00037-3
81. Zeng YJ, Ni YQ, Chen ZW, et al. (2025) Investigation of the impact of token embeddings in Transformer-based models on short-term tropical cyclone track and intensity predictions. *Eng Appl Comput Fluid Mech* 19: 2538180. https://doi.org/10.1080/19942060.2025.2538180
82. Bentsen LØ, Warakagoda ND, Stenbro R, et al. (2023) Spatio-temporal wind speed forecasting using graph networks and novel Transformer architectures. *Appl Energy* 333: 120565. https://doi.org/10.1016/j.apenergy.2022.120565
83. Zhou H, Zhang S, Peng J, et al. (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35: 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

AIMS Press