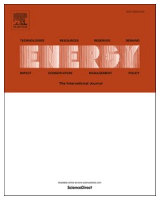
Energy 288 (2024) 129713

Contents lists available at ScienceDirect

Energy

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



Day-ahead wind power forecasting based on feature extraction integrating vertical layer wind characteristics in complex terrain

Keunmin Lee [a,b,*], Bongjoon Park [c,1], Jeongwon Kim [b], Jinkyu Hong [b]

[a] Digital Transformation (DX) Department, GS Windpower, Seoul, 06141, Republic of Korea
[b] Ecosystem-Atmosphere Process Laboratory, Department of Atmospheric Sciences, Yonsei University, Seoul, 03722, Republic of Korea
[c] Algorithm Lab, GS Energy, Seoul, 06141, Republic of Korea

ARTICLE INFO

Handling Editor: Jesse L. The

Keywords:
Wind power forecast
Complex terrain
Weather and research forecasting (WRF)
Light gradient boosting machine (LGBM)
Principal component analysis (PCA)

ABSTRACT

Accurate wind power forecasts help establish efficient power supply plans and stabilize power systems. For long-term forecasts, the outputs of numerical weather prediction (NWP) models are pipelined as inputs for the statistical post-processing model, underscoring the necessity of understanding forecasts simulated from NWP to enhance power prediction accuracy.

This study aims to enhance the quality of wind power forecasts in complex terrains, focusing on identifying and processing appropriate wind features of vertical layers simulated by NWP. In complex terrains with significant terrain variability, it is crucial to meticulously analyze and select the optimal vertical layer for each site or turbine individually, as simulated wind speeds at higher vertical layers than those used in previous research could potentially yield stronger correlations. Furthermore, we introduce a data processing technique that integrates wind characteristics across vertical layers, utilizing Principal Component Analysis (PCA). This approach not only provides physically intuitive results but also demonstrates enhanced performance compared to other feature selection methods. By selecting the appropriate vertical layers and applying the proper feature extraction, for wind farms situated in complex terrains in Korea, the annual normalized mean absolute error can be reduced by up to 1.2 %.

1. Introduction

Greenhouse gas emissions produced from the combustion of fossil fuels are significant contributors to this escalating climate change [1]. The IPCC suggests that renewable energy shows the greatest promise for mitigating net emissions, given its cost-effectiveness and high mitigation potential [2]. Among the renewable energy technologies currently under development, wind energy emerges as well-established and mature with substantial potential for large-scale expansion and advancement. Typically, areas near coastlines with high, consistent wind speeds and low levels of turbulence are ideal installation locations for large-scale wind farms, but unfortunately, such locations are often densely populated [3]. Offshore wind farms, alternatively, draw high installation costs and require resolving complexities in power grid integration. As such, due to the lack of better solutions, historically overlooked complex terrains with harsh wind conditions are now becoming more attractive to the wind power industry [4]. Wind turbines installed in these areas are primarily located on steep ridges with higher and stronger speed-up effects of the wind [5]. In South Korea, as of 2022, 62 % of the total 115 wind farms and 85 % of 33 wind farms exceeding 20 MW are situated in mountainous terrains [6]. As more and more wind energy facilities are placed in these terrains, there's a growing need for accurate wind prediction in complex terrains to ensure the stability of power system operations.

In long-term forecasts, it is well known that the performance of wind power forecast models depends on the terrain complexity of the wind farm, and as the complexity increases, forecasting errors become more and more dispersed [7,8]. For lead times of greater than 6 hour, it becomes imperative to leverage Numerical Weather Prediction (NWP) systems to accurately forecast wind and wind speeds [9]. However, mesoscale NWP models may produce uncertainties in the physical representation due to simplifications and parameterizations, struggling to comprehensively simulate the detailed characteristics of complex terrain [10,11]. Recently, high-resolution modeling and observational

* Corresponding author.
E-mail addresses: km.lee@yonsei.ac.kr, kmlee@gswind.com (K. Lee).
[1] Present addresses: Hyundai Mobis, Teheran-ro 203, Gangnam-gu, Seoul, 06141, Republic of Korea.

https://doi.org/10.1016/j.energy.2023.129713
Received 29 June 2023; Received in revised form 15 November 2023; Accepted 17 November 2023
Available online 22 November 2023

studies in complex terrains have been conducted to improve wind forecasts and the comprehension of boundary layer physics [12–17], yet even these high-resolution models have limitations in accurately capturing the microphysical process of real atmospheric motions and the sub-grid scale process. As a drawback of these NWP models, the simulated wind patterns at turbine height may not match the real wind conditions experienced by turbines, especially in complex terrain.

To overcome some of the limitations of NWP models and facilitate site-specific wind power predictions, statistical downscaling techniques of NWP outputs have been proposed [18]. Recent advancements have applied post-processing based machine learning methodologies, demonstrating greater prediction accuracy when compared to purely NWP-based forecasts [19,20], but most studies focus more on the development of statistical methods rather than exploring meaningful information provided by the NWP model [21]. However, intrinsically, the accuracy of wind power predictions, which leverage forecasted meteorological data from NWP models, is dependent on the accuracy of upstream forecasts. Thus, when selecting from the variety of meteorological data offered by NWP models for statistical post-processing algorithms, it is crucial to critically assess the input features, as recent studies have shown that careful selection can significantly improve the accuracy of wind speed and power predictions [20–22]. Salcedo et al. (2018) identified the 25 most suitable prediction features out of a set of 98 extracted for a single NWP grid point, enhancing the prediction accuracy in hourly and daily wind speeds by up to 20 % [22]. The chosen 25 input features include eastward and northward wind components from various vertical layers. Couto and Estanqueiro (2022) investigated 29 meteorological variables across multiple NWP spatial points targeting seven wind farms to achieve the best performance for each site. During this process, wind speeds at 50 and 110 m from various spatial grids were found to be crucial for predicting wind power across all wind farms [21]. Moreover, Gallego-Castillo et al. (2015) identified the most important features characterizing wind power variability during wind ramp events, including the eastward and northward wind components and the geopotential height at various pressure levels [23].

Even though previous studies have highlighted that one of the key features for predicting wind power is the forecasted wind information at various heights, studies exploring state-of-the-art statistical approaches still primarily utilize wind information from a single layer of the NWP model [24]. Some research considering wind information of various levels has limited their initial selection of vertical layers to those only within a physically intuitive range of 200 m (or solely the upper layer such as 500 hPa) and just explored these in a similar fashion to other exogenous features for the purpose of feature selection [21,22,25,26]. There has been limited investigation into which vertical layers should be used according to site-specific characteristics such as complex mountainous terrain.

Principal Component Analysis (PCA) is one of the most extensively utilized tools for feature extraction in predicting wind power production, with its applications predominantly concentrated on extracting features from only horizontal grids to capture the spatial variability of wind comprehensively [21,25–28]. In Ref. [21], PCA was conducted on each meteorological variable from horizontal grids surrounding the wind farm, yielding improved results compared to those achieved by selecting variables from a single point by integrating spatial patterns. However, wind within the Planetary Boundary Layer (PBL) is influenced not only by horizontal variability but also by vertical variability. In particular, in complex terrains, wind in vertical layers can be sensitive to this, as the PBL exhibits higher spatial variability compared to flat areas [29]. Therefore, we aimed to select the NWP vertical layers appropriate for complex site characteristics, and to assimilate the wind dynamics across various vertical layers through PCA, thereby enhancing the wind information used for predictions at the current given turbine altitude. To the best knowledge of the authors, no previous studies have examined the advantages of combining wind components using PCA feature extraction across multiple vertical layers in complex terrains.

With this background, the objective of this study is to identify and process wind components of vertical layers suitable for site-specific wind characteristics among supplementary variables simulated by NWP that can enhance the accuracy of deterministic wind power forecasting. This study analyzes the wind characteristics of a wind farm located in complex mountainous terrain and proposes a data processing method that integrates wind components across multiple vertical layers using PCA. In most cases, the feature extraction approaches are less interpretable than feature selection methods [22], yet our approach takes advantage of the orthogonal nature of the principal components to extract interpretable features that encapsulate the physical attributes of the northward and eastward wind components. The features extracted using the proposed technique retain not only information about wind speed and direction but also variance in the main wind fields across higher vertical layers. This results in enhanced correlations for the wind of each turbine, leading to additional predictive performance improvements. We assess the performance of the proposed method in a real-world application at a wind farm situated in the mountainous regions of Yeongyang-gun, South Korea, demonstrating improved performance in comparison to the existing approaches. Our research is noteworthy in that it explores a relatively limited wind power forecasting research in complex terrains covering wind characteristics and error characteristics of wind speed and wind power for individual turbines by conducting downscaling through mesoscale modeling, providing meteorologists with insights for more precise modeling.

1.1. Structure of this paper

The structure of this paper is as follows: Section 2 describes the proposed methodology integrating vertical layer wind components, and introduces details about the study site, the WRF model for numerical forecasts, the Light Gradient Boosting Machine (LGBM) to predict the wind speed and power of each turbine, and the details of the performed experiments. This study is based on a wind power forecast model that is currently being used in the field. Due to the requirement to submit accurate predictions within a limited time every day to Korea Power Exchange (KPX), which oversees the operation of Korea's electricity market and power system, we employ the LGBM model, which is known for its robustness, accuracy, and speed. Section 3 analyzes the wind speed and wind power characteristics of turbines situated in complex terrains and validates the improvement in prediction performance selecting the appropriate vertical layers with our proposed methodology. We conduct an error analysis on characteristics of wind speed and power considering the altitude of turbines. Section 4 concludes the paper with some key findings.

2. Methodology and data

2.1. The framework of proposed hybrid forecasting method

2.1.1. WRF simulation

Fig. 1 shows an outline of the proposed wind power forecasting method. The meteorological variables have been simulated from 2019 to 2022 for a total of 4 years with the WRF model in hindcast mode, using the Global Forecast System (GFS) dataset ran at 18:00 (UTC) operated by the National Center for Environmental Prediction (NCEP). The process involves a 3-h model spin-up with forecast times ranging from 18 to 45 h, enabling the forecasting from 00:00 to 24:00 (KST) of the subsequent day.

The WRF (v.4.1.3) model is configured with two domains arranged as one-way nests, centered on the Yeongyang wind farm, with each domain containing 30 × 30 grids (Fig. 2a). The horizontal resolutions of these domains are tripled, ranging from 9 to 3 km. The vertical height limit is set to 50 hPa and 33 eta layers are used. The geographic information in the WRF model comes from 30-s resolution data provided by the United States Geological Survey (USGS). The average surface

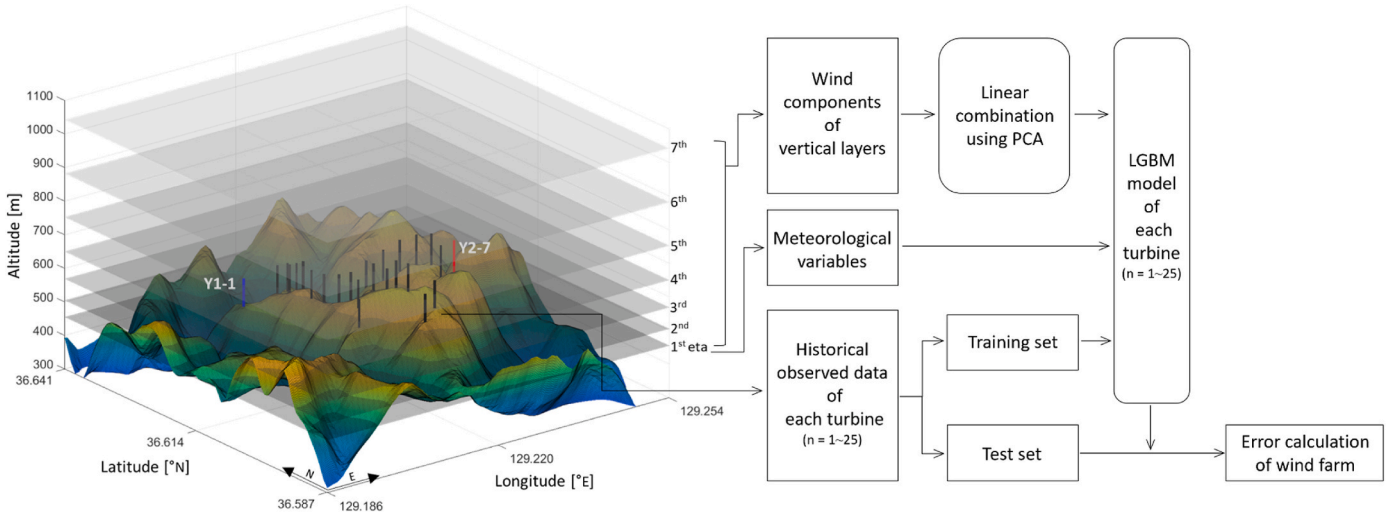


Fig. 1. Diagram of the proposed wind power forecasting method.

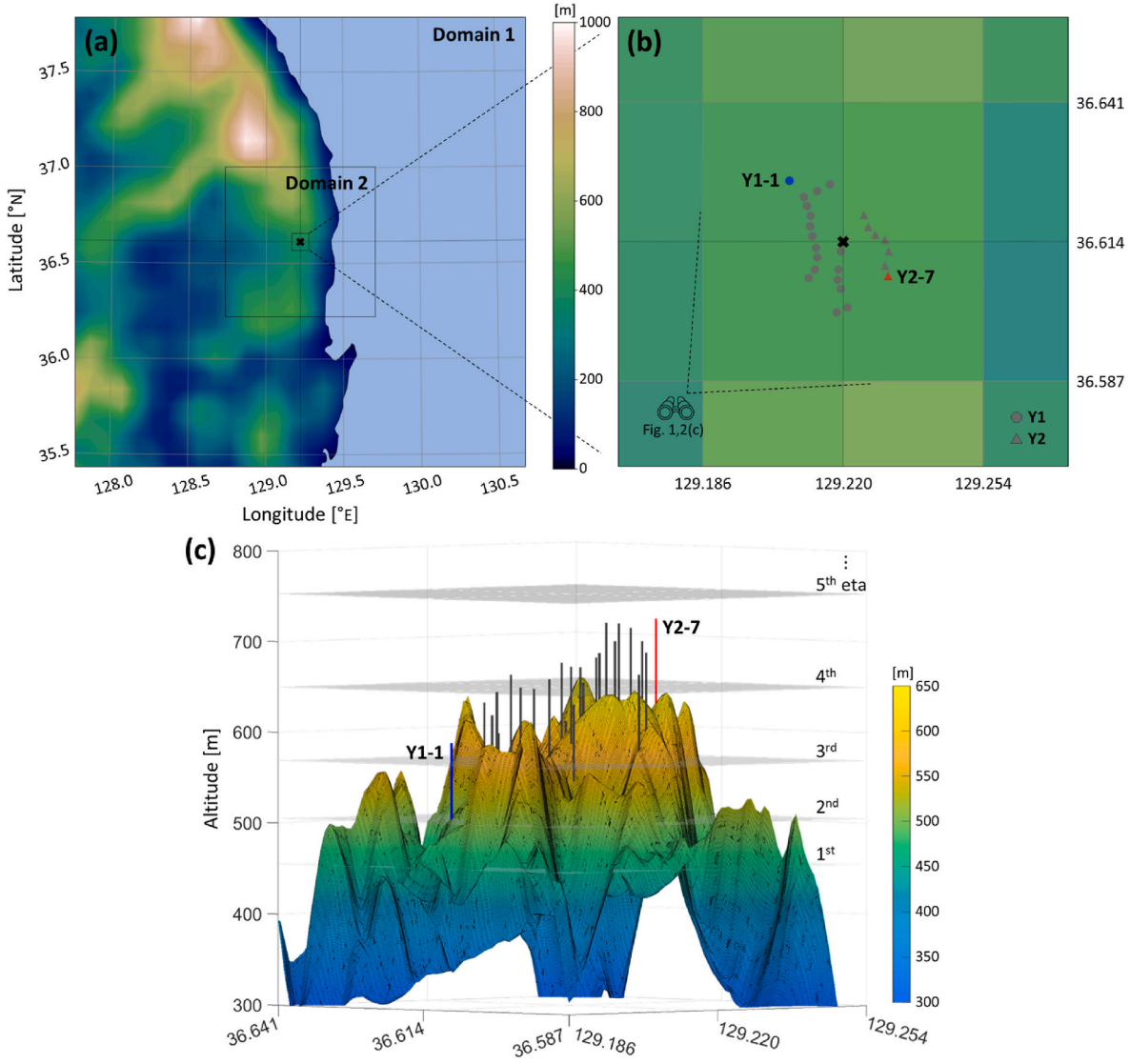


Fig. 2. (a) The domain setting of the WRF model and the surface elevation height of USGS 30s. (b) The enlarged domain 2 with the distribution of 25 turbines of Yeongyang wind farm. 'X' indicates the central point of the wind farm. (c) The actual altitude of 25 turbines as implemented with the 90 m resolution DEM of NGII, and the vertical layer setting for the WRF model.

elevation of the four adjacent grids in domain 2 where the wind farm is located is 454.8 m (Fig. 2b).

The meteorological variables are obtained by averaging four adjacent grids in domain 2. Similarly, the eastward and northward wind components (u, v) are averaged across the two staggered grids within the neighboring four grids, with the purpose of finding the value at the central point (36.614 °N, 129.22 °E). The meteorological variables at a height of 80 m above ground level are calculated from the 1st eta layer, at which the wind components are derived by the Monin-Obukhov similarity theory.

The u, v wind components from eta layers near the ground are used for PCA to obtain information about the representative wind field at turbines located in the complex terrain. Consequently, a total of 7 eta layers, ordered from closest to the surface to the farthest, are chosen out of 13 through comparative experiments of prediction performance, with accompanied analysis of the observed wind speed (Section 3.2). The heights from the 1st to the 13th eta layer are 0, 50.0, 113.9, 195.2, 298.0, 427.2, 587.8, 785.5, 1025.6, 1312.6, 1649.6, 2037.7, and 2475.6 m from the ground (454.8 m) respectively.

2.1.2. Light Gradient Boosting Machine

The subsequent post-processing step involves using LGBM in parallel for each of the 25 turbines to train the relationship between historically observed data of each turbine and the WRF simulated next day (forecast time; 21–45 h) data. While the majority of existing studies adhere to wind farm level predictions, we adopted a point-wise prediction method for each turbine and then aggregated the results to estimate the power of the wind farm, due to the distinct wind characteristics of individual turbines at our study site [30].

Gradient Boosted Trees (GBT) algorithms [31,32], a large family of machine learning techniques encompassing LGBM, have been used in recent research for wind power forecasting [25,33]. In GBT, the n+1st tree sums the product of the nth prediction error and the learning rate (lr) to iteratively minimize the error (Eq. 1–2). The final predicted value $\hat{y}$ is represented by adding the prediction for the error associated with the number of boosting rounds (n_br) to the average value of the training set ($\bar{y}_{tr}$) (Eq. (3)).

$$\hat{y}_{n+1} = \hat{y}_n + lr \bullet f_{n+1}(x) \tag{1}$$

$$f_{n+1} = \underset{f}{\operatorname{argmin}}\, Loss(y - \hat{y}_n, f(x)) \tag{2}$$

$$\hat{y} = \bar{y}_{tr} + lr \bullet \sum_{i=1}^{n\_br} f_i(x) \tag{3}$$

LGBM utilizes two key techniques, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), to enhance both prediction speed and performance in ensemble decision tree-based algorithms [34]. GOSS optimizes memory usage by preserving data points with large gradient magnitudes, essential for maximizing information gain, while selectively utilizing only a subset of data points with smaller gradient magnitudes. EFB focuses on feature engineering by grouping together various features and identifying and eliminating unnecessary ones, thus reducing dimensionality. One distinctive feature of LGBM is its use of histograms for discretizing continuous features into bins during tree construction. This histogram-based algorithm reduces computational complexity and conserves memory, contributing to its efficiency. Another key characteristic is the leaf-wise growth strategy, which expands trees horizontally, helping to avoid unnecessary memory usage during tree construction [34]. Consequently, LGBM is renowned for its exceptional speed and accuracy, especially when dealing with extensive high-dimensional datasets or imbalanced data distributions [35]. In comparative experiments conducted with Multilayer Perceptron (MLP) and Gated Recurrent Unit (GRU), LGBM exhibited superior performance to both, showing robustness (i.e., strong results without in-depth hyperparameter tuning) especially when dealing with a large number of features as indicated in Fig. S1 in the supplementary material. In terms of speed, LGBM demonstrated results 15 times faster than MLP and 32 times faster than GRU. The benefit of decreased computational time allows us to experiment with various combinations of inputs and more effectively correlate the influence of factors on observed outcomes.

In this study, the model is trained with 10,000 iterations and a learning rate of 0.05, with early stopping rounds set at 50 to prevent overfitting. The decision trees have a maximum depth of 10 to control model complexity. Bagging and feature fractions are both set to 0.8, introducing randomness for robustness. The model's performance is evaluated using the Root Mean Square Error (RMSE) metric to measure prediction accuracy.

2.2. Site description and historical observed data

The Yeongyang wind farm is comprised of two plants, Y1 and Y2, which consist of 25 wind turbines manufactured by Vestas with a total capacity of 83.6 MW. Y1 has 18 wind turbines of 3300 kWh with a rotor diameter of 114 m and a hub height of 84 m, while Y2 has 7 wind turbines of 3450 kWh with a rotor diameter of 114 m and a hub height of 94 m, a slightly higher tower height and output. The 25 turbines are located within a radius of 1.7 km from the center of the site (Fig. 2b). Table 1 presents the latitude, longitude, and altitude of each turbine. In this study area, the ruggedness index (RIX) value is 42, based on the Digital Elevation Model (DEM) from the National Geographic Information Institute (NGII). The RIX is defined as the percentage of terrain that has a slope greater than a particular threshold, 0.3 in this case, within a 3500 m radius of the center of the site [36]. Within the WAsP system, this metric is used to assess topographical complexity: a flat area exhibits a RIX of 0, whereas a notably steep region can reach values close to 30 [37]. Thus, this study area exhibits significant topographical complexity, with 42 % of its landscape exceeding the critical slope. The average (± standard deviation) elevation of the generator hub of the 25 turbines is 664.5 ± 38.7 m. Even within the same wind farm, there is a significant difference in hub height of approximately 140 m between the lowest (Y1-1) and the highest (Y2-7) turbines (Fig. 2c). The wind speed data is measured using two 2D-Ultrasonic Anemometer (FT-702-LT)

Table 1
Location, altitude, and hub height of 25 turbines in Yeongyang wind farm.

Turbine	Latitude [°]	Longitude [°]	Altitude [m]	Hub height [m]
Y1-1	**36.626**	**129.207**	**502.7**	
Y1-2	36.623	129.211	547.5	
Y1-3	36.624	129.214	532.6	
Y1-4	36.625	129.217	510.8	
Y1-5	36.621	129.211	560.4	
Y1-6	36.619	129.212	578.6	
Y1-7	36.617	129.212	565.0	
Y1-8	36.615	129.213	564.3	
Y1-9	36.613	129.213	576.1	
Y1-10	36.611	129.214	593.7	84
Y1-11	36.612	129.219	586.6	
Y1-12	36.609	129.213	589.6	
Y1-13	36.607	129.212	548.7	
Y1-14	36.609	129.219	598.0	
Y1-15	36.607	129.219	638.3	
Y1-16	36.605	129.219	637.1	
Y1-17	36.601	129.221	605.9	
Y1-18	36.600	129.218	581.5	
Y2-1	36.619	129.225	515.5	
Y2-2	36.617	129.226	559.2	
Y2-3	36.615	129.228	590.6	
Y2-4	36.614	129.230	604.7	94
Y2-5	36.612	129.231	619.0	
Y2-6	36.609	129.230	605.3	
Y2-7	**36.607**	**129.231**	**630.8**	
Average	**36.614**	**129.220**	**577.7**	

installed on the nacelle behind the blades. Wind speed and wind power generation data from 25 turbines have been collected by the supervisory control and data acquisition (SCADA) system for 4 years from 2019 to 2022.

2.3. Principal components analysis of vertical layer wind components

In the NWP model, the wind components u and v represent the eastward and northward wind vectors, respectively. The following equations are commonly used to convert u and v into wind speed (ws) and wind direction (wd) to provide precise information about wind characteristics to machine learning applications:

$$ws = \sqrt{u^2 + v^2} \tag{4}$$

$$wd = 270 - atan2(v, u) \tag{5}$$

Therefore, when using wind information from multiple vertical layers, there arises a challenge of dealing with a growing number of input features for u, v, ws, and wd for each layer.

PCA is a mathematical technique used in multivariate analysis to transform high-dimensional data into a low-dimensional space, increasing interpretability while minimizing information loss [38]. PCA is commonly used for reducing dimensions by identifying new features through linear combinations of features of the original dataset, capturing as much variance of the data as possible. By transforming existing features into uncorrelated principal components (*PCs*) that maximize variance, PCA retains most of the significant information from the original data space. The first *PC* represents the direction of maximum variance, and subsequent components are orthogonal to prior components.

The fundamental steps of PCA are as follows:

1) Let X be a $n \times 2m$ matrix of the wind component data, where n is the number of total available forecasts for the study, and m is the number of vertical layers. The matrix of wind components from multiple vertical layers is first defined as

$$X = [\, X_1 \quad \cdots \quad X_{2m} \,] = \begin{bmatrix} x_{1,1} & \cdots & x_{1,2m} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,2m} \end{bmatrix} = [\, U_1 \quad \cdots \quad U_m \quad V_1 \quad \cdots \quad V_m \,] = \begin{bmatrix} u_{1,1} & \cdots & u_{1,m} & v_{1,1} & \cdots & v_{1,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{n,1} & \cdots & u_{n,m} & v_{n,1} & \cdots & v_{n,m} \end{bmatrix} \tag{6}$$

where m = 7 in this study (Section 3.2). X_1, ⋯, X_m correspond to the u components from vertical layers, and X_{m+1}, ⋯, X_{2m} correspond to the v components from vertical layers.

2) Then, the covariance matrix C of X matrix has a symmetric $2m \times 2m$ matrix. The eigenvector V is obtained by applying eigendecomposition to the covariance matrix C.

3) The *PCs* are calculated from the dot product of the matrix X and eigenvector V. In general, *PCs* are defined as X^*V where the i,j $^{\text{th}}$ sample data of column-centered matrix X^* computed as $x_{i,j}{}^* = x_{i,j} - \bar{x}_j$, for facilitating a direct connection to a geometric approach of PCA [38]. In this case, X^* becomes a rotational transformation around the column-center, $\bar{x}_j$, and the origin of transformed *PCs* no longer represent a windless physical state. Also, considering that u and v wind components are vectors signifying the west − east and north − south directions, respectively, it results in a loss of the physical interpretation of the wind direction for *PCs* as they are transformed through linear combinations of the column-centered u and v wind components. Consequently, the *PCs* calculated by performing rotational transformation around the zero point through the dot product between X without subtracting the mean value of column and V have consistent units (m s^{-1}) that encompass both ws and wd information.

4) Given that the eigenvalues of C be $\lambda_1, \cdots, \lambda_{2m}$, are the proportion of variance explained of the p$^{\text{th}}$ *PC*, E_p, then the cumulative proportion of variance explained from the 1$^{\text{st}}$ to p$^{\text{th}}$ *PCs*, $E(p)$ is calculated as:

$$E_p = \frac{\lambda_p}{\sum_{k=1}^{2m} \lambda_k} \tag{7}$$

$$E(p) = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{2m} \lambda_k} \tag{8}$$

2.4. Experimental setting

2.4.1. Experimental design factoring for wind information in vertical layers

The ranges of the two sets of experimental data are 1 year (test set; from January 1, 2022, to December 31, 2022), and 3 years (training and validation set at a 2:1 ratio; from January 1, 2019, to December 31, 2021). These periods were selected as historical observation data have been measured for verification consistently for this duration.

Table 2 shows the configuration of experiments to validate the improvement in predictive performance considering the impact of the wind information of the model's vertical layers. Based on previous studies, the most commonly used meteorological variables simulated at the turbine height are selected for the control experiment (CTL). At turbine height (80 m) from the ground, 7 basic meteorological variables are used as input features: temperature (t), specific humidity (q), wind speed (ws), wind direction (wd), eastward and northward wind vectors (u, v), and surface pressure ($psfc$). In the first experimental condition (EXP_1), wind information (u_1, v_1, ws_1, wd_1) from the first vertical layer is used as input features to assess the improvement in performance from adding wind information from vertical layers. In subsequent experiments, in the EXP_2 to EXP_{13}, more and more layers are included, in the n_{th} experiment EXP_n, wind information ($u_{1\sim n}$, $v_{1\sim n}$, $ws_{1\sim n}$, $wd_{1\sim n}$) up to the n_{th} vertical layer is added to iteratively assess the predictive per-

Table 2
Details of experiments for wind features of vertical layers.

Experiments	Input features	Number of features	Target features
CTL	10 min basic meteorological variables on the next day (21–45 h) simulated by WRF. At 80 m, temperature (t), specific humidity (q), wind speed (ws), direction (wd), eastward and northward wind vectors (u, v), and surface pressure ($psfc$).	7	Historical wind speed or wind power observed at each turbine at 10 min
EXP_1	CTL + ws_1, wd_1, u_1, v_1	11	
	Add wind component up to the corresponding vertical layer to the CTL.		
EXP_2	CTL + $ws_{1\sim2}$, $wd_{1\sim2}$, $u_{1\sim2}$, $v_{1\sim2}$	15	
⋮	⋮	⋮	
EXP_7	CTL + $ws_{1\sim7}$, $wd_{1\sim7}$, $u_{1\sim7}$, $v_{1\sim7}$	35	
⋮	⋮	⋮	
EXP_{13}	CTL + $ws_{1\sim13}$, $wd_{1\sim13}$, $u_{1\sim13}$, $v_{1\sim13}$	59	

Table 3
Details of experiments for feature selection and extraction approach.

Experiments	Input features	Methodology
FS_{CORR}	28 wind features of seven vertical layers in EXP_7	Feature selection (filter method based linear correlation coefficient)
FS_{RE}		Feature selection (wrapper method based recursive elimination)
FE_{AE}		Feature extraction (Autoencoder)
FE_{PCA}		Feature extraction (PCA)

formance improvement of the addition of each vertical layer.

2.4.2. Experimental design for feature selection and extraction approach

The comparative experiments are designed to appropriately process a large number of wind features ($u_{1\sim7}$, $v_{1\sim7}$, $ws_{1\sim7}$, $wd_{1\sim7}$) from a total of 7 vertical layers (Table 3), focusing on feature selection and extraction. Filter methods are preprocessing strategies that employ quality metrics, derived directly from an examination of the data [39]. Wrapper methods utilize learning algorithms to identify the subset of features that provide the highest predictive performance, and LGBM is used for this purpose [40].

In the FS_{CORR} approach, we filter for wind features from 3 vertical layers with high linear correlations between observed and forecasted *ws* for each turbine. This selection process is conducted independently of the predictive model. In the FS_{RE} approach, a wrapper-based feature selection method is employed, recursively eliminating wind features that contributed the least to performance improvement based on feature importance. The performance is evaluated using validation errors, and the set of features leading to the best predictive performance are chosen for each turbine. In the FE_{AE} approach, feature extraction of the wind features is performed with autoencoders: neural network structures that encode input data into a low-dimensional latent space and decode it back to the original data [41]. The compressed latent features obtained through the encoder represent a new dimensionally reduced input vector while preserving data features. In this experiment, an autoencoder with two layers is used to encode wind features into 8 latent spaces, which are then added as input features to the LGBM model. Finally, in the FE_{PCA} approach, the proposed PCA method is applied to extract 8 *PCs*, which are used as input features.

2.5. Statistical criteria

The RMSE, bias, variance, and Pearson correlation coefficient (r) are defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i)^2} \tag{9}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i) \tag{10}$$

$$variance = \frac{1}{n-1}\sum_{i=1}^{n}[(pred_i - obs_i) - \text{bias}]^2 \tag{11}$$

$$r = \frac{\sum_{i=1}^{n}(pred_i - \overline{pred})(obs_i - \overline{obs})}{\sqrt{\sum_{i=1}^{n}(pred_i - \overline{pred})^2 \sum_{i=1}^{n}(obs_i - \overline{obs})^2}} \tag{12}$$

where *pred* represents the simulated variable, *obs* represents the observation of each turbine, and n represents the number of hourly data. The Pearson correlation coefficient quantifies the linear correlation between observed and forecasted data, with values between −1 and 1. A value of +1 indicates a perfect positive linear correlation, −1 is a perfect negative correlation, and values near zero suggest little to no correlation. In forecasting, a value near zero indicates a weak forecast, while a value close to 1 signifies an accurate prediction and a high negative value indicates phase opposition.

In this study, the Normalized Mean Absolute Error (NMAE; %) is used to evaluate forecasts against observations, adhering to the guidelines set by the KPX. The NMAE for the wind farm is defined as:

$$NMAE = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{turbine} pred_i - \sum_{turbine} obs_i\right|}{\sum_{turbine} installed\ capacity} \times 100 \tag{13}$$

where the installed capacity is 3300 kWh for Y1 turbines and 3450 kWh for Y2 turbines. When calculating the error of the turbine, NMAE divides the error of each wind turbine by its installed capacity, but when calculating the error of the entire wind farm, the total installed capacity of the wind farm of 83,550 kWh (3300 kWh × 18 + 3450 kWh × 7) is used.

This post-processing study employs data with 10-min resolutions to forecast wind speed and power. However, given the common practice of forecasting the renewable energy sector on an hourly basis, statistical values for error calculation are presented as hourly intervals.

3. Results and discussion

3.1. Wind characteristics of turbines in complex terrain

Wind speed (*ws*) varies significantly within the wind farm in complex mountainous terrain depending on the altitude and topography of where the turbines are installed (Fig. 3a). The annual average (± standard deviation) *ws* for the lowest (Y1-1) and highest turbine (Y2-7) among the 25 turbines are 4.8 (± 2.6) and 7.1 (± 3.6) m s^{-1} respectively (Y2-7 about 1.5 times stronger than Y1-1). Similarly, the annual average wind power generations for Y1-1 and Y2-7 are 471.0 (± 669.2) and 1138.8 (± 1158.2) kWh respectively, showing a difference of about 667.8 kWh (about 2.4 times larger for Y2-7). The wind power shows a clear tendency to increase as the *ws* increases with altitude, demonstrating a proportional relationship between the average *ws* and wind powers of each turbine (Fig. 3). At these higher altitudes, the *ws* and wind power is greater, but is accompanied by larger variation, leading to greater prediction errors.

In the wind farm, the *ws* is stronger in winter and weaker in the summer (Fig. 4a). December is the windiest month, with an average *ws* of 8.0 m s^{-1}, while July is the least with 4.5 m s^{-1}. The Y2-7 turbine experiences an average *ws* of over 10 m s^{-1} in the early morning of December, while the Y1-1 turbine has an average *ws* of less than 6 m s^{-1}. Also, different patterns of monthly diurnal variation in *ws* are observed for individual turbines. During the nighttime, there is an increased discrepancy among the turbines, which tends to decrease throughout the daytime. Specifically, turbines located at higher altitudes, where wind conditions are stronger, experience a decrease in *ws* from dawn to noon. On the other hand, turbines situated at lower altitudes exhibit weaker winds in the early morning, which gradually intensify after sunrise. This discrepancy in *ws* patterns can be attributed to the formation of the nocturnal boundary layer within PBL at night, as well as the subsequent development of the mixed layer after sunrise. During the nighttime, turbines positioned at higher altitudes experience stronger winds due to the presence of the nocturnal boundary layer where *ws* increases with altitude. However, after sunrise, the mixed layer develops, and the differences in *ws* between different altitudes tend to decrease [29]. This highlights that there is a substantial difference in diurnal patterns of *ws* depending on the topographical characteristics of the turbine, even among turbines within the same wind farm.

Fig. 4b illustrates the monthly data distribution categorized by wind direction and time averaged from 25 turbines. The wind direction is generally similar for all turbines located close together within a radius of 1.7 km, with a predominant westward wind observed throughout most

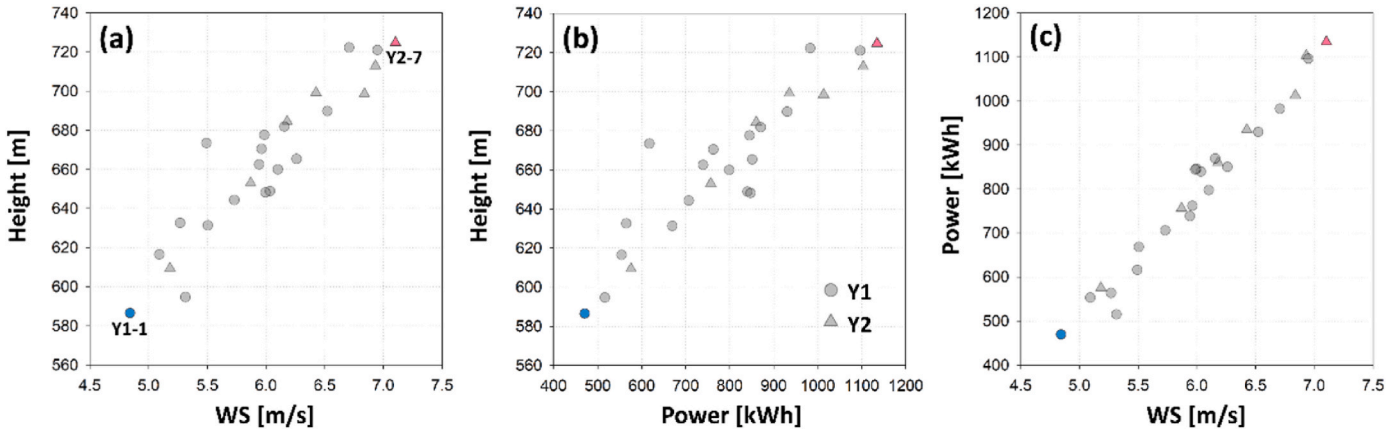


Fig. 3. For 25 turbines, (a) relationship between observed wind speed and height of wind turbine hubs, (b) relationship between observed wind power and height of wind turbine hubs, (c) relationship between observed wind speed and power.

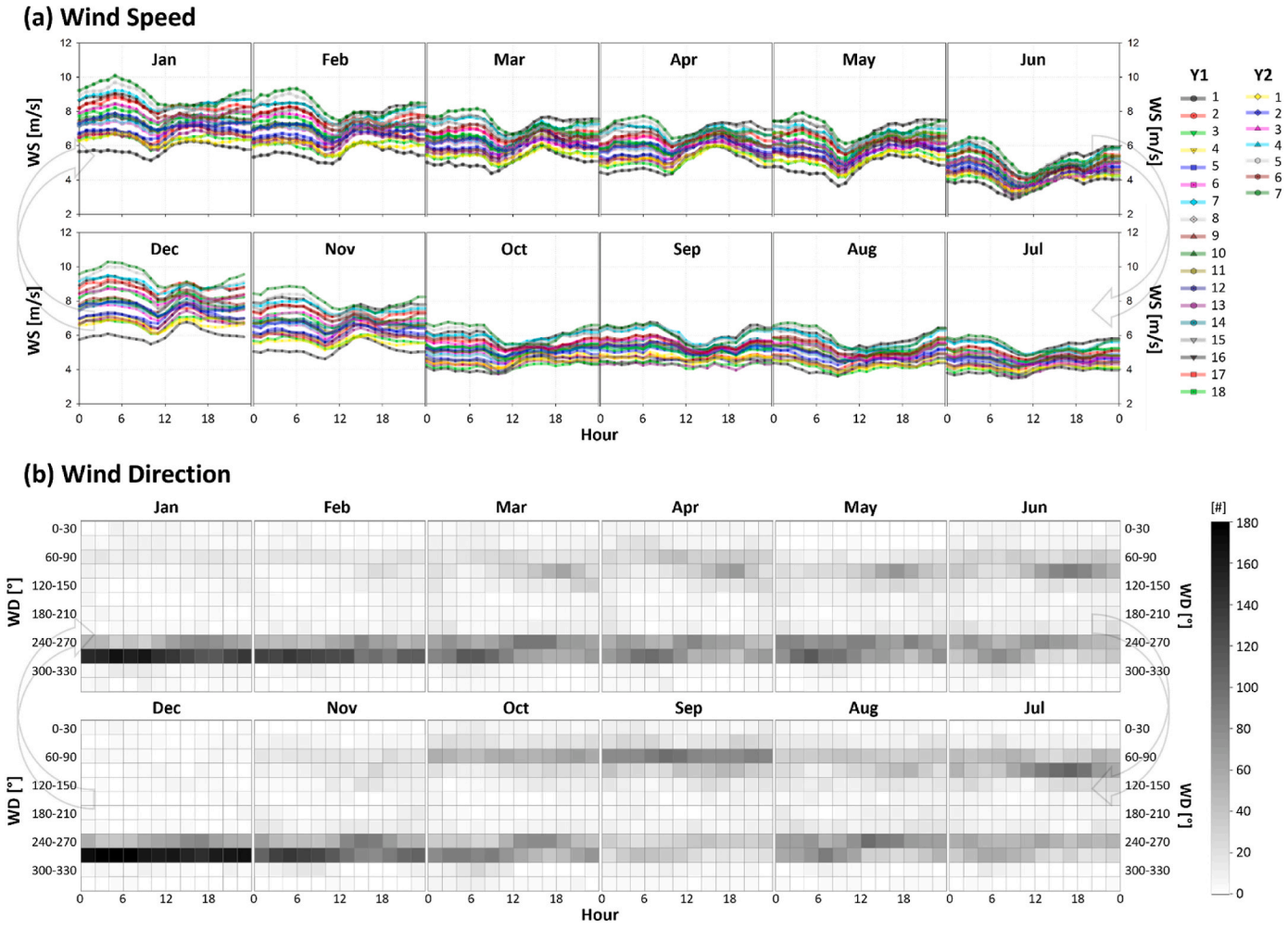


Fig. 4. For 4 years from 2019 to 2022, (a) monthly mean diurnal change in observed wind speed of the 25 turbines, (b) the number of data partitioned by wind direction and time of each month. The partitions of wind direction and time are indicated every 30° and 2 h.

seasons. This westward direction is particularly pronounced during the winter. However, in June and July, there is a noticeable change in wind direction corresponding to sunrise and sunset, attributed to the influence of mountain valley winds from the warming and cooling of the mountain ridges. Additionally, during the fall season in September and October, the wind direction can vary depending on synoptic conditions, exhibiting eastward wind direction.

3.2. *Determining the number of vertical layers to match wind characteristics of turbines in complex terrain*

Fig. 5 indicates the correlation between *ws* forecasted at different

	(m)	(hpa)	Avg	Y1-1	Y1-2	Y1-3	Y1-4	Y1-5	Y1-6	Y1-7	Y1-8	Y1-9	Y1-10	Y1-11	Y1-12	Y1-13	Y1-14	Y1-15	Y1-16	Y1-17	Y1-18	Y2-1	Y2-2	Y2-3	Y2-4	Y2-5	Y2-6	Y2-7
eta_1	0	946	**0.756**	0.729	0.745	0.738	0.733	0.750	0.760	0.760	0.764	0.780	0.778	0.758	0.755	0.730	0.750	0.740	0.762	0.760	0.730	0.736	0.773	0.789	0.762	0.773	0.770	0.778
eta_2	50	940	**0.763**	0.742	0.752	0.743	0.742	0.759	0.770	0.767	0.770	0.788	0.788	0.767	0.761	0.736	0.761	0.754	0.775	0.762	0.733	**0.741**	0.777	0.796	0.770	0.778	**0.770**	0.782
eta_3	114	932	**0.767**	0.747	0.756	0.745	0.746	0.762	0.776	0.770	0.773	0.792	0.794	0.772	0.765	0.739	0.768	0.763	0.785	0.762	0.739	0.740	0.777	0.797	0.775	0.780	0.769	0.785
eta_4	195	922	**0.770**	0.749	0.759	0.747	0.750	0.764	0.780	0.772	0.775	0.794	0.798	0.776	0.768	0.743	0.774	0.770	0.793	0.762	0.744	0.739	0.777	**0.797**	0.780	**0.783**	0.769	**0.788**
eta_5	298	910	**0.773**	0.752	0.764	0.751	**0.752**	0.768	0.784	0.777	0.778	**0.795**	**0.801**	**0.779**	0.771	0.749	0.778	0.775	0.798	**0.763**	0.749	0.739	**0.777**	0.796	**0.785**	0.783	0.769	0.788
eta_6	427	894	**0.772**	**0.754**	0.769	**0.753**	0.751	0.773	**0.786**	**0.781**	**0.781**	0.794	0.800	0.776	**0.773**	**0.753**	**0.779**	**0.776**	**0.798**	0.760	**0.751**	0.736	0.773	0.789	0.783	0.774	0.763	0.780
eta_7	588	876	**0.763**	0.750	**0.770**	0.748	0.740	**0.773**	0.781	0.779	0.778	0.785	0.791	0.765	0.769	0.752	0.772	0.768	0.789	0.748	0.744	0.725	0.757	0.771	0.767	0.752	0.744	0.759
eta_8	786	854	0.741	0.733	0.760	0.731	0.716	0.762	0.764	0.765	0.761	0.764	0.772	0.740	0.755	0.741	0.754	0.750	0.766	0.722	0.727	0.701	0.727	0.735	0.735	0.714	0.711	0.723
eta_9	1026	828	0.703	0.700	0.735	0.699	0.675	0.734	0.731	0.733	0.728	0.726	0.737	0.701	0.727	0.717	0.722	0.717	0.727	0.682	0.696	0.660	0.679	0.682	0.685	0.659	0.661	0.672
eta_{10}	1313	798	0.653	0.655	0.694	0.653	0.623	0.691	0.684	0.687	0.681	0.676	0.690	0.651	0.684	0.677	0.677	0.670	0.675	0.630	0.650	0.608	0.622	0.622	0.623	0.597	0.601	0.612
eta_{11}	1650	765	0.598	0.602	0.644	0.601	0.567	0.640	0.629	0.633	0.626	0.620	0.635	0.596	0.633	0.630	0.623	0.616	0.619	0.577	0.599	0.552	0.562	0.561	0.560	0.537	0.542	0.553
eta_{12}	2038	729	0.544	0.545	0.590	0.549	0.514	0.583	0.572	0.576	0.568	0.564	0.578	0.541	0.581	0.580	0.564	0.557	0.560	0.529	0.550	0.501	0.508	0.507	0.502	0.486	0.491	0.502
eta_{13}	2476	690	0.497	0.494	0.539	0.504	0.469	0.531	0.521	0.525	0.518	0.516	0.529	0.494	0.534	0.535	0.509	0.502	0.508	0.492	0.509	0.458	0.463	0.465	0.455	0.447	0.452	0.463

Fig. 5. Each turbine's correlation between observed wind speeds and forecasted wind speeds at different vertical layers is represented, with higher correlations indicated by green colors and lower correlations by red. Colors are distinguished for each turbine, and the altitude value (m) denotes the height above the average surface elevation of the four adjacent grids in domain 2, which is 454.8 m. The layers with the highest correlation for each turbine are highlighted in bold. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

vertical layers for grid points and observed *ws* of 25 turbines within the Yeongyang wind farm. At 11 turbines among the 25 turbines, the 6th vertical layer exhibits the highest correlation with *ws*, followed by the 5th vertical layer at 7 sites, the 4th vertical layer at 3 sites, the 7th vertical layer and the 2nd vertical layer at 2 sites. When considering the average values across all 25 sites, the 5th and 6th vertical layers (approximately 300–430 m above ground level) showed the highest correlation in *ws*, followed by the 4th, 3rd, 7th, and 2nd vertical layers in that order. Above the 8th vertical layer, the correlation gradually decreases across all 25 turbines. The heights exhibiting the highest correlation, 5th and 6th vertical layers, are beyond the range of the ground-to-200 m (up to 3rd or 4th layer) used in previous studies [21,22,25,26,42]. This could be due to the fact that the spatial variability of wind in complex mountainous terrain is higher compared to flat terrain. Secondarily, the ground information used in mesoscale modeling is simply an averaged value across the entire grid, which could lead to relative inaccuracies. In this study, for instance, the average elevation of the four nearby grids of domain 2 of 30-s resolution USGS geographic data is 454.8 m, while the average actual elevation of the 25 turbine locations is 577.7 m, resulting in an approximate 120 m difference (Fig. 2c, Table 1). These findings underscore the importance of selecting precise vertical layers through analysis when choosing input features for post-processing models in regions with significant terrain variability, such as mountainous terrain. Additionally, this emphasizes the need for a finer vertical resolution with a reasonable horizontal resolution that effectively captures the terrain features in complex terrain to reduce the *ws* error simulated by the numerical forecasting model [10,11,15].

In the case of the CTL experiment, where only simulated meteorological data at turbine height is used, the wind power NMAE for the entire wind farm is 9.77 (Fig. 6 and Table S2). When sequentially adding wind features from lower to higher altitudes, a significant improvement is observed [21,22], when utilizing wind information up to 7th vertical layer with a high correlation to observed *ws* (Fig. 5). Subsequently, when utilizing wind information from 8th vertical layers, and onwards there is a tendency for performance improvement to saturate. Moreover, adding more than 11th upper-level layers results in a decline in predictive performance.

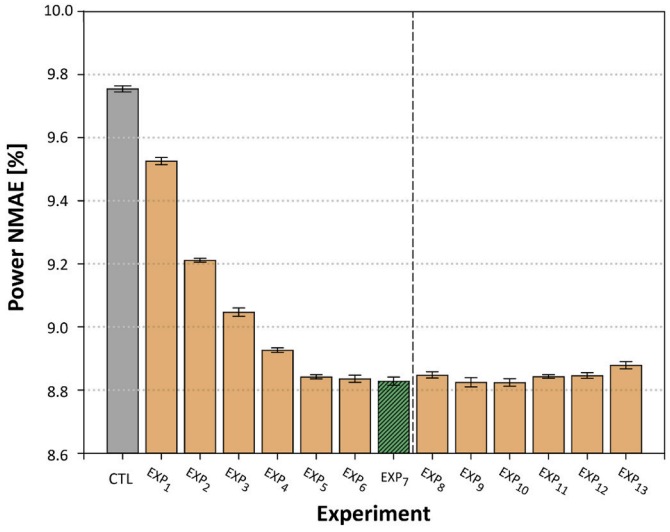


Fig. 6. Predictive performance when cumulatively adding wind information for each vertical layer. The NMAE represents the value for the entire wind farm. The error bars indicate the standard deviation from 10 repeated predictions. A more detailed box plot comparing with other machine learning methods can be found in the supplementary material (Fig. S1).

3.3. Characteristics of PCs of vertical layer wind fields

Table 4 displays the eigenvectors for 8 *PCs* of the 14 WRF-simulated wind features (*u*, *v* component of 7 layers). The eigenvectors are coefficients that describe the relationship between the *PCs* and the original variables. The equation for obtaining the eigenvectors of the covariance matrix is sign-invariant, and only the relative magnitudes and sign patterns of the eigenvectors are relevant [38]. By analyzing the eigenvectors, it is evident that the first *PC* (pc_1), mainly reflects the u component of multiple layers, while the second *PC* (pc_2) mainly reflects the v component. By the inherent attributes of the PCA methodology, which identifies succeeding components orthogonal to the primary *PC*, the two *PCs* excel at transcribing the properties of the orthogonal *u* and *v* vectors. By assessing the absolute values of the eigenvectors corresponding to each level, one can ascertain that the vertical levels which are more prominently represented in the *PCs* contain more significant wind information. In particular, the u components of 5th to 7th and v components from 4th to 6th vertical layers stand out, indicating that a greater amount of information from the higher layers contributes to pc_1 and pc_2 when compared to the layers closer to the surface.

The pc_1 accounts for 79.5 % of the total variance, while pc_2 accounts for 17.6 % (Fig. S3). This argues the significance of winds in the x direction at the wind farm, where westerly winds are dominating (Fig. 4b). The cumulative variance contributed by the pc_1 and pc_2 amounts to 97.1 %. It is noteworthy that the dominant wind encompassing information

Table 4
Eigenvectors for each principal component of the WRF simulated 14 variables (u, v components of 7 layers). The pc_1 and pc_2 direction is indicated based on the wind direction.

	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8	pc_1 direction	pc_2 direction
u_1	**−0.29**	0.00	0.02	−0.37	0.19	−0.49	0.05	0.55	95.7 (275.7)	1.1 (181.1)
v_1	0.03	**−0.25**	−0.42	−0.01	−0.55	−0.20	−0.51	0.05		
u_2	**−0.35**	−0.02	0.01	−0.39	0.06	−0.29	−0.01	−0.14	95.6 (275.6)	3.0 (183.0)
v_2	0.03	**−0.33**	−0.43	−0.03	−0.23	−0.09	0.25	−0.01		
u_3	**−0.37**	−0.03	0.00	−0.32	−0.03	0.05	−0.01	−0.46	96.1 (276.1)	4.5 (184.5)
v_3	0.04	**−0.38**	−0.33	−0.01	0.14	0.05	0.46	−0.01		
u_4	**−0.39**	−0.04	−0.02	−0.18	−0.10	0.37	−0.02	−0.27	97.0 (277.0)	6.3 (186.3)
v_4	0.05	**−0.41**	−0.15	0.00	0.41	0.11	0.11	0.00		
u_5	**−0.40**	−0.06	−0.02	0.05	−0.11	0.47	−0.03	0.28	98.3 (278.3)	8.5 (188.5)
v_5	0.06	**−0.42**	0.10	−0.02	0.39	0.07	−0.40	−0.03		
u_6	**−0.40**	−0.08	0.01	0.35	−0.03	0.18	−0.04	0.45	99.8 (279.8)	11.6 (191.6)
v_6	0.07	**−0.41**	0.37	−0.05	0.04	−0.01	−0.37	−0.03		
u_7	**−0.40**	−0.10	0.09	0.66	0.06	−0.46	0.06	−0.33	101.3 (281.3)	15.2 (195.2)
v_7	0.08	**−0.38**	0.60	−0.09	−0.48	−0.05	0.38	0.06		

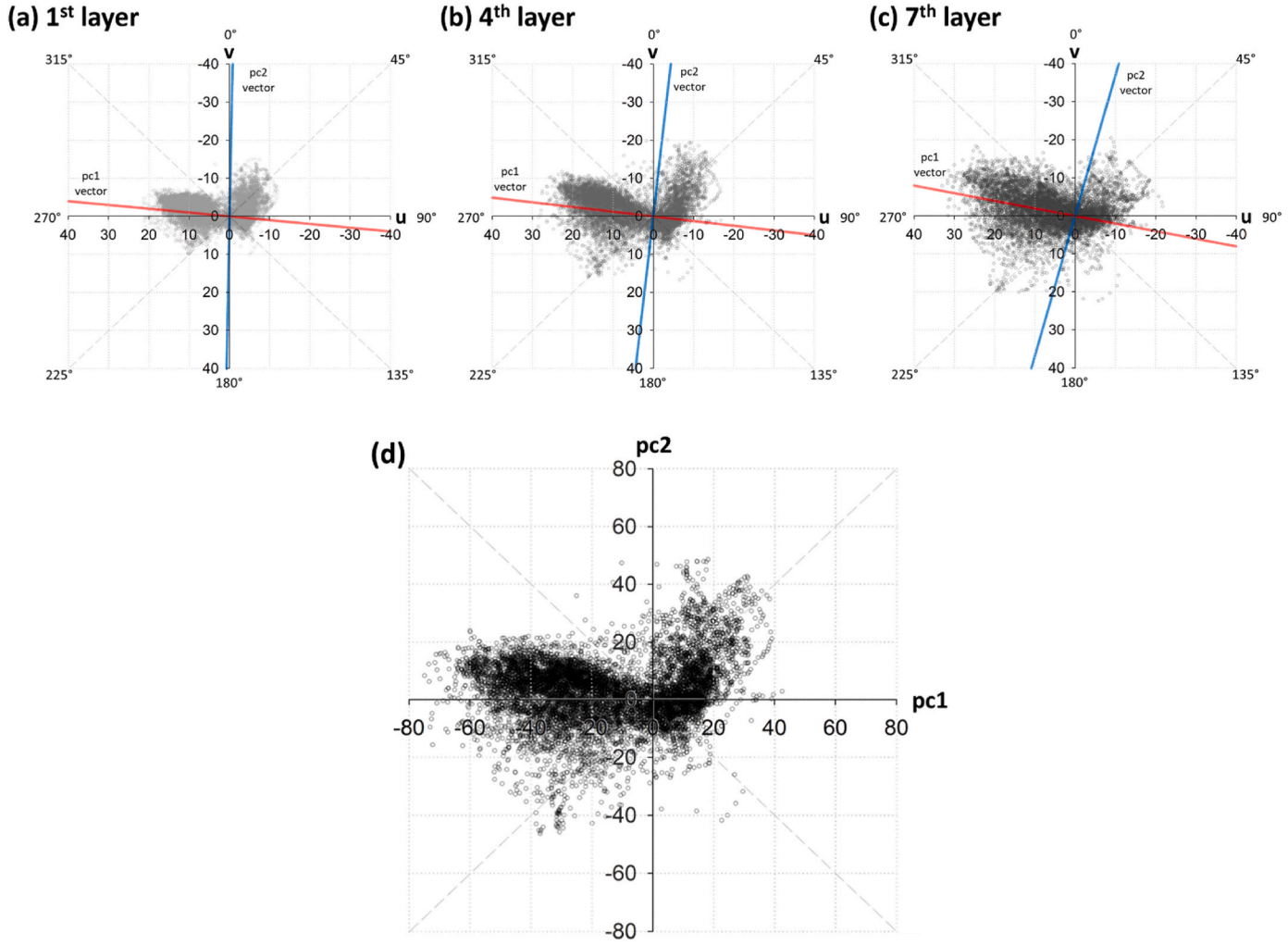


Fig. 7. Wind components of (a) 1st, (b) 4th, and (c) 7th vertical layers with pc_1 (red) and pc_2 (blue) vectors projected onto each vertical layer. In order to indicate the wind direction by sector, the signs of the u and v axes are reversed. (d) pc_1, pc_2 which is a linear combination of the u, v components of the seven vertical layers. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

from multiple vertical layers can be effectively represented by using only two variables. As wind information across vertical layers is highly correlated, PCA effectively eliminates interdependent information and isolates distinct, variable-specific data. While pc_1 and pc_2 capture a significant portion of the information, the addition of more *PCs* yields only a slight improvement in prediction accuracy. This study utilizes up to 8th *PCs* to enhance the accuracy of wind power predictions (Tables 3 and 4).

Fig. 7a-c shows the wind components of the 1st, 4th, and 7th vertical layers with pc_1 and pc_2 vectors projected onto each layer. The pc_1 vector

corresponds to the west − east direction, which aligns with the dominant wind direction, whereas the pc_2 vector denotes the south − north direction. The direction of pc_1 and pc_2 vectors in each layer veers clockwise with height, exhibiting a clockwise rotation relative to the 1st layer (Table 4). This rotation is revealed to be the Ekman spiral pattern, characterized by a clockwise rotation of the wind pattern as the frictional force diminishes further away from the surface. Fig. 7d reveals the distribution of pc_1 and pc_2 which is a linear combination of the u, v components of the seven vertical layers. The pc_1 value encompassing the majority of variance information related to the primary wind field in the region, is larger than that of pc_2 (Fig. S3).

An examination of the time series of absolute pc_1 values and ws across each vertical layer shows that pc_1 adequately accounts for ws, explaining approximately 80 % of wind variance in regions where the predominant wind direction is west − east (Fig. 8a). However, in periods marked by the prevalence of northerly and southerly winds, such as June 6–7, 14, and 28 − 29, the ws representation is not fully encompassed by pc_1 alone (yellow part in Fig. 8). This representation is considerably improved through the inclusion of the pc_2 variable (Fig. 8b), and it is evident that $\sqrt{pc_1 + pc_2}$, l_2-norm of pc_1 and pc_2 expressed in the form of ws (Eq. 4), tends to follow the patterns of the 4th and 7th layers rather than the 1st layer, due to PCs reflecting more information from the 4th to 7th vertical layers (Table 4). Moreover, the pc_1 variable, which mirrors the u component, changes sign approximately at 180°, effectively illustrating changes in wind direction (Fig. 8c and d). The wind direction features exhibit unstable and inconsistent changes, displaying discontinuities at 0 and 360°, which can complicate the machine learning training process for wind direction data. This issue can be mitigated if PCs can provide a more concise representation of wind speed and direction across various vertical layers.

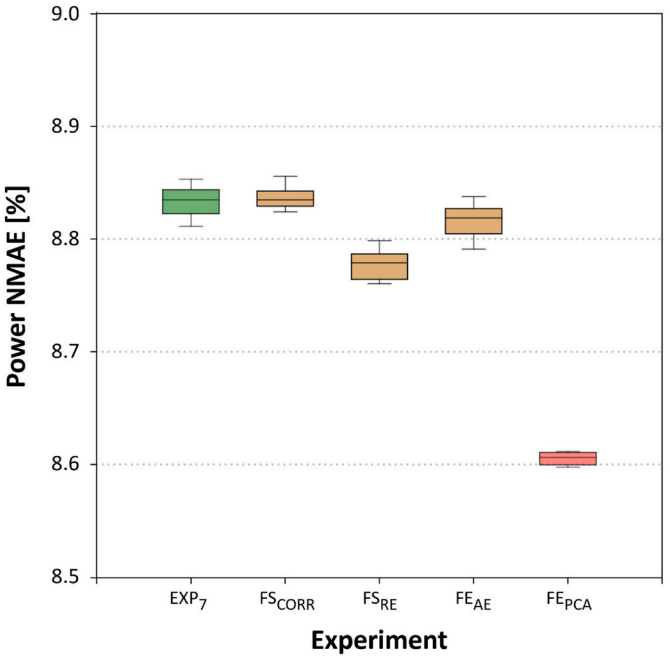


Fig. 9. Box plot comparison of wind power prediction accuracy (NMAE), across ten iteration predictions, before and after applying feature selection and extraction methods to wind features from seven vertical layers in EXP_7.

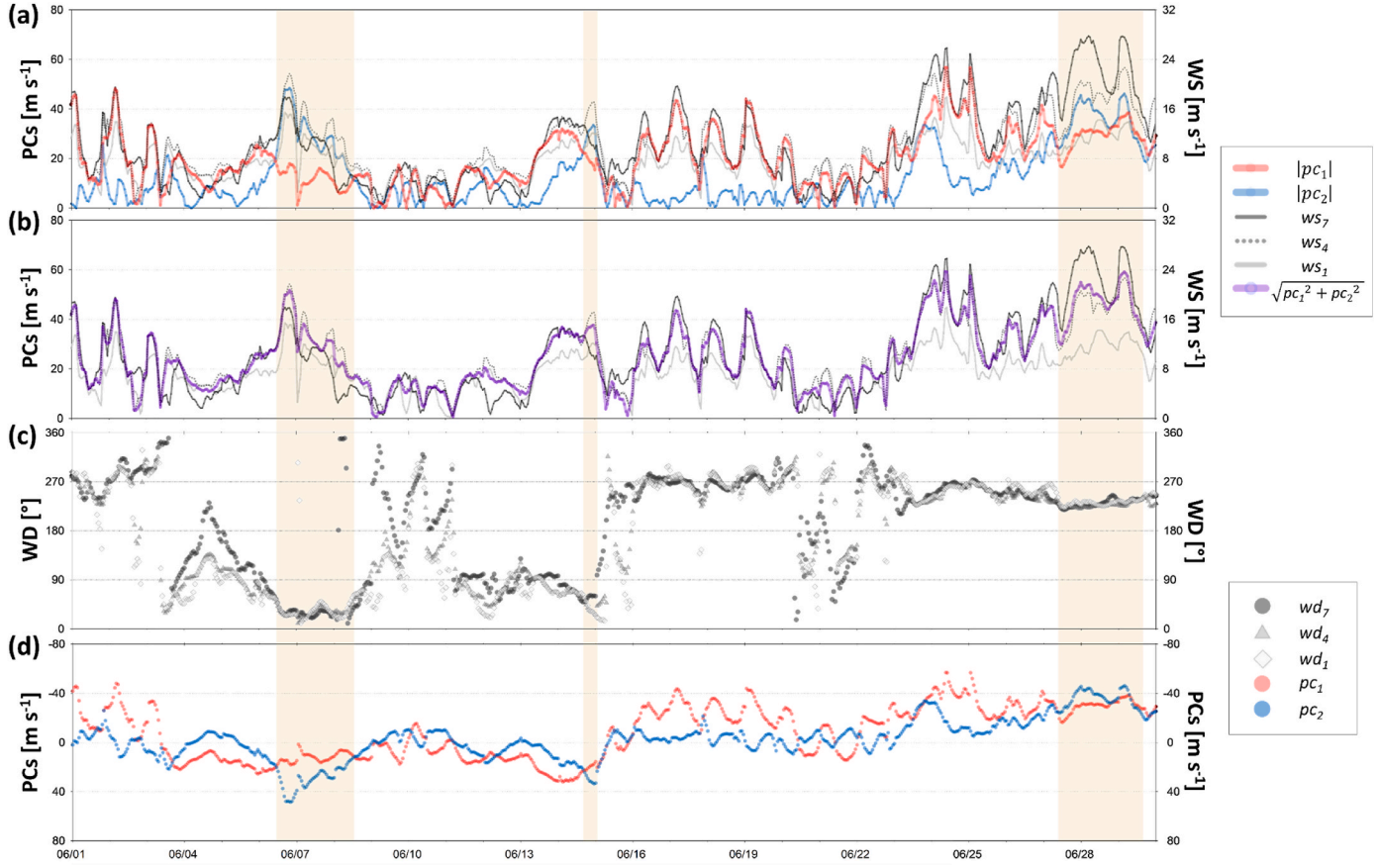


Fig. 8. In June 2022, (a) the absolute value of pc_1 and pc_2, (b) the l_2-norm of pc_1 and pc_2 with wind speed for 1st, 4th, and 7th vertical layers, and (c) wind direction for 1st, 4th, and 7th vertical layers, (d) pc_1, pc_2.

3.4. Understanding the benefit of using PCs of vertical layer wind fields

We compare the wind power prediction results using wind features from 7 vertical layers (EXP_7) before and after applying feature selection and extraction methods to wind features (Table 3, Fig. 9). Through the results of the FS_{CORR} experiment, it is found that an average NMAE stand at 8.82 by using only wind features from 3 vertical layers with the highest correlation with observed *ws* per turbine. This is identical to the performance of EXP_7, which results in an average NMAE of 8.82. For the FS_{RE} experiment, which recursively removed features based on their feature importance in the validation set, the average NMAE improves to 8.77.

In the FE_{AE} results, predictions using 8 features extracted through an autoencoder without considering their physical meaning for each vertical layer show no significant performance improvement compared to EXP_7. This suggests that simply applying feature extraction for dimensionality reduction may not yield substantial performance enhancements. In the case of FE_{PCA}, which uses 8 extracted *PCs*, there is a significant improvement, with an average NMAE of 8.61. Table 5 presents the results of the linear correlation between observed *ws* and $\sqrt{pc_1 + pc_2}$ being derived with the extracted pc_1 and pc_2, which capture the meanings of *u* and *v* components. For all turbines, the correlation of extracted *PCs* by integrating vertical layer wind characteristics exhibits better results than that of the forecasted *ws* that have the highest correlation among each vertical layer. This indicates that wind information enhanced by integrating multiple vertical layer wind characteristics through PCs, which places greater importance on 4th to 7th layers with high correlation (Fig. 5), leads to improved performance.

Previous studies utilized the wind features of vertical layers in a similar fashion to other exogenous features for the purpose of feature selection to find the optimal combination for the best prediction performance [21,22], but these results suggest the necessity of feature extraction with physical meaning, especially for turbines located in complex terrain that are sensitive to the influence by vertical variability of PBL (Fig. 4). Additionally, the feature extraction using PCA has the advantage of being more than 10 times faster than the wrapper-based feature selection. The proposed technique can also serve as an efficient way to reduce the excessive increase in input features when considering the wind features across multiple vertical layers.

Table 5
Comparison of observed wind speed correlations between both the highest correlated wind speed among vertical layers and the l_2-norm of pc_1 and pc_2 for 25 turbines.

Turbine	Correlation of *ws* (at vertical layer with the best correlation)	Correlation of $\sqrt{pc_1 + pc_2}$
Y1-1	0.754 (6th)	0.767
Y1-2	0.770 (7th)	0.780
Y1-3	0.753 (6th)	0.765
Y1-4	0.752 (5th)	0.764
Y1-5	0.773 (7th)	0.785
Y1-6	0.786 (6th)	0.798
Y1-7	0.781 (6th)	0.793
Y1-8	0.781 (6th)	0.795
Y1-9	0.795 (5th)	0.810
Y1-10	0.801 (5th)	0.814
Y1-11	0.779 (5th)	0.790
Y1-12	0.773 (6th)	0.786
Y1-13	0.753 (6th)	0.764
Y1-14	0.779 (6th)	0.790
Y1-15	0.776 (6th)	0.786
Y1-16	0.798 (6th)	0.808
Y1-17	0.763 (5th)	0.777
Y1-18	0.751 (6th)	0.761
Y2-1	0.741 (2nd)	0.754
Y2-2	0.777 (5th)	0.792
Y2-3	0.797 (4th)	0.810
Y2-4	0.785 (5th)	0.795
Y2-5	0.783 (4th)	0.794
Y2-6	0.770 (2nd)	0.783
Y2-7	0.788 (4th)	0.800
Average	**0.774**	**0.787**

3.5. Improvement of forecast performance through PCs of vertical layer wind fields

The average bias of post-processed *ws* across the 25 turbines is 0.03 m s^{-1} and 0.02 m s^{-1} in the CTL and FE_{PCA} experiments, respectively (Table S2). In both experimental conditions, all turbines demonstrate consistent near-zero bias independent of the altitude of the turbines (Fig. 10a). This indicates that the parallel configuration of machine learning-based post-processing uniquely configured for each turbine effectively mitigates systematic errors by considering the specific characteristics of each turbine [30].

The average variance of post-processing *ws* of the 25 turbines is reduced to 2.38 m^2 s^{-2} in the FE_{PCA} compared to 2.84 m^2 s^{-2} in the CTL (Table S2). The FE_{PCA} exhibits a pronounced improvement in performance across all turbines (Fig. 10b), demonstrating that *PCs* that integrate wind characteristics of 7 vertical layers improve the interpretation of the variance of wind across all turbines. There is a distinct pattern of variance of post-processed *ws* increasing with altitude, implying that the stronger the *ws* at higher altitudes, the greater the challenges posed for accurate prediction (Fig. 3). The NMAE of wind power also exhibits an enhanced performance for all turbines in the FE_{PCA} relative to the CTL. The improvement for each turbine is approximately 1.2 % (± 0.3 %). Exhibiting a similar increase with altitude as seen in the variance for post-processed *ws*, the NMAE of wind power ranges from 6.8 % for the lowest Y1-1 turbine to 12.0 % for the highest Y2-7 turbine (Fig. 10c). The equation for the NMAE involves offsetting of the forecast errors of each turbine (Eq. 13), and as such the NMAE of the wind farm is lower than the average NMAE for the 25 turbines (stars in Fig. 10c) aggregating the output from different wind turbines can be beneficial in reducing the prediction error [30]. Even among adjacent locations and at similar heights (ex. Y2-2 and Y2-4 at about 700 m height), within the same wind farm, the *ws* correlation is different (Fig. 5) and there is a large difference in NMAE. This significant discrepancy of NMAE can be the different wind characteristics of each turbine arising from their different topographical characteristics. Therefore, to achieve accurate wind power predictions in complex terrains, it is imperative to establish a model with a finer vertical resolution with a reasonable horizontal resolution that reflects accurate geographic information [10,11,15].

4. Conclusions and perspectives

In this study, wind information from various vertical layers of a

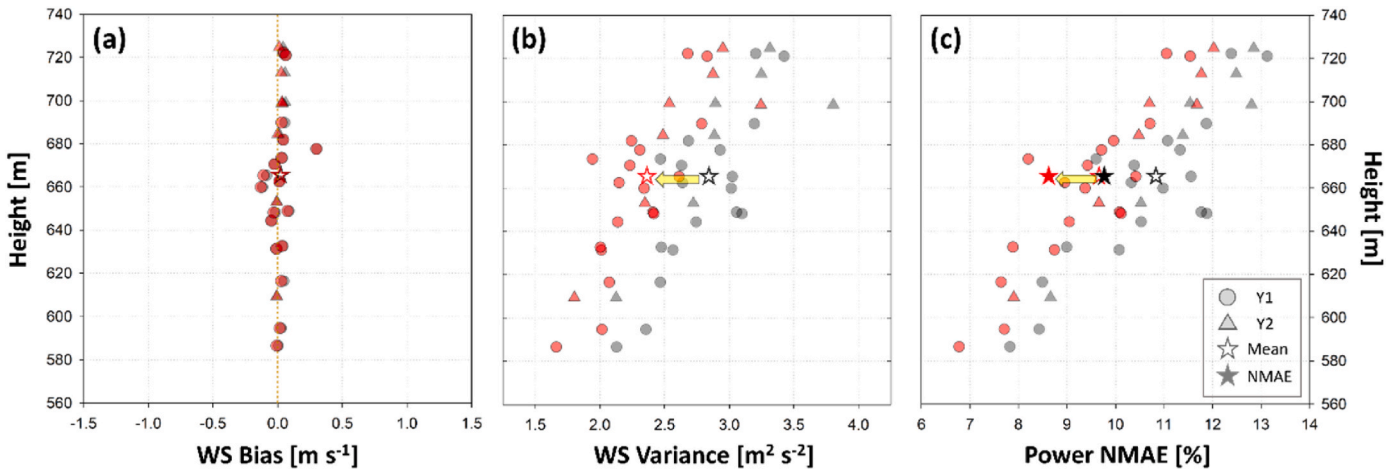


Fig. 10. During the test period, (a) the bias and (b) the variance of the post-processing wind speed and (c) the NMAE of the wind power of CTL (grey) and FE_{PCA} (red) experiments for 25 turbines by hub height. The blank stars indicate that the values averaging NMAE of 25 turbines, and the filled stars indicate NMAE of the wind farm. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

numerical weather prediction model is explored to enhance the accuracy of wind power forecasting in complex terrain. We aim to improve predictive performance by selecting appropriate wind features of vertical layers simulated by NWP for each site or turbine and integrating wind component information from multiple vertical layers through feature extraction. The WRF model is conducted over the complex mountainous region of Yeongyang, South Korea, and the LGBM model is utilized as the statistical model for wind power prediction.

In the case of wind farms located in mountainous complex terrain, there is a substantial difference in diurnal patterns of wind speed depending on the topographical characteristics of the turbine, even among turbines within the same wind farm. In complex terrain, the highest correlation between observed and forecast wind speeds was demonstrated at higher layers (approximately 300–430 m above ground level), beyond the range of vertical layers used in previous studies, necessitating the selection of the optimal vertical layers for each site or turbine.

The features extracted considering the physical characteristics of the vertical layer wind components enhance the correlation with wind speed, by retaining not only information about wind speed and direction but also variance in the main wind fields across higher layers above the turbine. Our approach exhibits improvement in performance compared to methods that involve selecting a subset of features through feature selection or feature extraction methods that do not take physical meaning into consideration. When the optimal vertical layers are selected and the proper feature extraction methodology is applied, the average annual NMAE is reduced by 1.2 % compared to using vertical layers only up to turbine height.

Even when post-processing is conducted for each turbine individually, while it helps mitigate systematic errors in wind prediction, there are still observed differences in variance even among turbines located at similar altitudes. This discrepancy can be attributed to the different wind characteristics of each turbine arising from their different topographical characteristics. Therefore, in future research aimed at improving the performance of wind power prediction in complex terrain, we plan to perform NWP modeling with high-resolution that accurately reflects geographic information and more detailed vertical layers near the surface. We will explore research on horizontal and vertical variability of wind to further enhance prediction accuracy and different model architectures, such as deep learning alternatives. Additionally, the methodology we propose may be effective for complex terrains that are sensitive to winds at higher levels, but the wind features extracted through integrating multiple layers do not necessarily represent the optimal combination for each vertical layer. Therefore, in future research, we intend to conduct studies to optimize the combination of wind components at different vertical layers to best suit the wind conditions for each turbine.

CRediT authorship contribution statement

Keunmin Lee: Conceptualization, Methodology, Software, Formal analysis, Data curation, Supervision, Project administration, Visualization, Writing - original draft, Writing - review & editing. **Bongjoon Park:** Conceptualization, Methodology, Software, Validation, Writing - review & editing. **Jeongwon Kim:** Methodology. **Jinkyu Hong:** Conceptualization, Resources, Funding acquisition.

Declaration of competing 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.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors are grateful to GS Windpower for supporting this research. The authors are grateful to Mr. Ho Seok Lee (Schwartz Lab, University Health Network, Toronto, Canada) for his support in proofreading the manuscript and interpreting PCA results. This study was conducted with the support of the Korea Meteorological Administration Research and Development Program (KMI2021-01610) and the Air Quality Forecasting Center at the National Institute of Environmental Research under the Ministry of Environment (NIER-2023-04-02-052).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2023.129713.

References

[1] Masson-Delmotte VP, Zhai P, Pirani SL, Connors C, Péan S, Berger N, et al., IPCC. Summary for policymakers. In: Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change. Cambridge, United Kingdom and New

York, NY, USA: Cambridge University Press; 2021. p. 3–32. https://doi.org/10.1017/9781009157896.001.
[2] Shukla PR, Skea J, Reisinger A, Slade R, Fradera R, Pathak M, et al., IPCC. Summary for policymakers. In: Climate change 2022: mitigation of climate change. Contribution of working group III to the sixth assessment report of the intergovernmental panel on climate change. Cambridge, UK and New York, NY, USA: Cambridge University Press; 2022. https://doi.org/10.1017/9781009157926.001.
[3] Alfredsson PH, Segalini A. Introduction Wind farms in complex terrains: an introduction. Phil Trans Math Phys Eng Sci 2017;375(2091):20160096.
[4] Elgendi M, AlMallahi M, Abdelkhalig A, Selim MY. A review of wind turbines in complex terrain. Int J Thermofluids 2023:100289.
[5] Lange J, Mann J, Berg J, Parvu D, Kilpatrick R, Costache A, et al. For wind turbines in complex terrain, the devil is in the detail. Environ Res Lett 2017;12(9):094020.
[6] KWEIA. Annual report on wind energy industry in Korea. Korea Wind Power Industry Association; 2022. 2023.
[7] Kariniotakis G, Martí I, Casas D, Pinson P, Nielsen TS, Madsen H, et al. What performance can be expected by short-term wind power prediction models depending on site characteristics? European Wind Energy Conference EWEC; 2004.
[8] Marti I, Kariniotakis G, Pinson P, Sanchez I, Nielsen TS, Madsen H, et al. Evaluation of advanced wind power forecasting models–results of the ANEMOS Project. In: European wind energy conference. EWEC; 2006. p. 9.
[9] Soman SS, Zareipour H, Malik O, Mandal P. A review of wind power and wind speed forecasting methods with different time horizons. North Am Power Symposium 2010:1–8.
[10] Horvath K, Koracin D, Vellore R, Jiang J, Belu R. Sub-kilometer dynamical downscaling of near-surface winds in complex terrain using WRF and MM5 mesoscale models. J Geophys Res Atmos 2012;117(D11).
[11] Giebel G, Badger J, Perez IM, Louka P, Kallos G, Palomares AM, et al. Shortterm forecasting using advanced physical modelling-the results of the anemos project. In: Proceedings of the European wind energy conference; 2006 February.
[12] Prósper MA, Otero-Casal C, Fernández FC, Miguez-Macho G. Wind power forecasting for a real onshore wind farm on complex terrain using WRF high resolution simulations. Renew Energy 2019;135:674–86.
[13] Shaw WJ, Berg LK, Cline J, Draxl C, Djalalova I, Grimit EP, et al. The second wind forecast improvement project (WFIP2): general overview. Bull Am Meteorol Soc 2019;100(9):1687–99.
[14] Wilczak JM, Stoelinga M, Berg LK, Sharp J, Draxl C, McCaffrey K, et al. The second wind forecast improvement project (WFIP2): observational field campaign. Bull Am Meteorol Soc 2019;100(9):1701–23.
[15] Olson JB, Kenyon JS, Djalalova I, Bianco L, Turner DD, Pichugina Y, et al. Improving wind energy forecasting through numerical weather prediction model development. Bull Am Meteorol Soc 2019;100(11):2201–20.
[16] Bianco L, Muradyan P, Djalalova I, Wilczak JM, Olson JB, Kenyon JS, et al. Comparison of observations and predictions of daytime planetary-boundary-layer heights and surface meteorological variables in the columbia river gorge and basin during the second wind forecast improvement project. Boundary-Layer Meteorol 2022;182(1):147–72.
[17] Radünz WC, de Almeida E, Gutiérrez A, Acevedo OC, Sakagami Y, Petry AP, et al. Nocturnal jets over wind farms in complex terrain. Appl Energy 2022;314:118959.
[18] Hanifi S, Liu X, Lin Z, Lotfian S. A critical review of wind power forecasting methods—past, present and future. Energies 2020;13(15):3764.
[19] Liu H, Chen C, Lv X, Wu X, Liu M. Deterministic wind energy forecasting: a review of intelligent predictors and auxiliary methods. Energy Convers Manag 2019;195:328–45.
[20] Liu H, Chen C. Data processing strategies in wind energy forecasting models and applications: a comprehensive review. Appl Energy 2019;249:392–408.
[21] Couto A, Estanqueiro A. Enhancing wind power forecast accuracy using the weather research and forecasting numerical model-based features and artificial neuronal networks. Renew Energy 2022;201:1076–85.
[22] Salcedo-Sanz S, Cornejo-Bueno L, Prieto L, Paredes D, García-Herrera R. Feature selection in machine learning prediction systems for renewable energy applications. Renew Sustain Energy Rev 2018;90:728–41.
[23] Gallego-Castillo C, Garcia-Bustamante E, Cuerva A, Navarro J. Identifying wind power ramp causes from multivariate datasets: a methodological proposal and its application to reanalysis data. IET Renew Power Gener 2015;9(8):867–75.
[24] Khazaei S, Ehsan M, Soleymani S, Mohammadnezhad-Shourkaei H. A high-accuracy hybrid method for short-term wind power forecasting. Energy 2022;238:122020.
[25] Andrade JR, Bessa RJ. Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Trans Sustain Energy 2017;8(4):1571–80.
[26] Salcedo-Sanz S, Pastor-Sánchez A, Prieto L, Blanco-Aguilera A, García-Herrera R. Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization–Extreme learning machine approach. Energy Convers Manag 2014;87:10–8.
[27] Kusiak A, Zheng H, Song Z. Wind farm power prediction: a data-mining approach. Wind Energy: An Int J for Progress and Applic Wind Power Convers Technol 2009;12(3):275–93.
[28] Davò F, Alessandrini S, Sperati S, Delle Monache L, Airoldi D, Vespucci MT. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. Sol Energy 2016;134:327–38.
[29] Stull RB. An introduction to boundary layer meteorology, vol. 13. Springer Science & Business Media; 1988.
[30] Yakoub G, Mathew S, Leal J. Intelligent estimation of wind farm performance with direct and indirect 'point'forecasting approaches integrating several NWP models. Energy 2023;263:125893.
[31] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–232.
[32] Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197–227.
[33] Singh U, Rizwan M, Alaraj M, Alsaidan I. A machine learning-based gradient boosting regression approach for wind power production forecasting: a step towards smart grid environments. Energies 2021;14(16):5196.
[34] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.
[35] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67.
[36] Wood N. The onset of separation in neutral, turbulent flow over hills. Boundary-Layer Meteorol 1995;76(1–2):137–64.
[37] Bowen AJ, Mortensen NG. WAsP prediction errors due to site orography. Riso National Laboratory; 2004. p. 28–9.
[38] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Phil Trans Math Phys Eng Sci 2016;374(2065):20150202.
[39] Carta JA, Cabrera P, Matías JM, Castellano F. Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study. Appl Energy 2015;158:490–507.
[40] Feng C, Cui M, Hodge BM, Zhang J. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. Appl Energy 2017;190:1245–57.
[41] Bengio Y, Courville AC, Vincent P. Unsupervised feature learning and deep learning: a review and new perspectives. CoRR 2012;5538(2665):1. abs/1206.
[42] Lu P, Ye L, Zhao Y, Dai B, Pei M, Li Z. Feature extraction of meteorological factors for wind power prediction based on variable weight combined method. Renew Energy 2021;179:1925–39.