Renewable and Sustainable Energy Reviews xxx (xxxx) xxx

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

An overview of performance evaluation metrics for short-term statistical wind power forecasting

J.M. González-Sopeña [a], V. Pakrashi [b,c,*], B. Ghosh [a]

[a] Department of Civil, Structural and Environmental Engineering, Museum Building, Trinity College Dublin, Dublin 2, Ireland
[b] Dynamical Systems and Risk Laboratory, School of Mechanical & Materials Engineering, University College Dublin, Dublin, Ireland
[c] SFI MaREI Centre, University College Dublin and the Energy Institute, University College Dublin, Ireland

ARTICLE INFO

Keywords:
Wind power forecasting
Accuracy estimation
Performance evaluation metrics
Hybrid decomposition-based models

ABSTRACT

Wind power forecasting has become an essential tool for energy trading and the operation of the grid due to the increasing importance of wind energy. Therefore, estimating the forecast accuracy of a WPF model and understanding how the accuracy is calculated are necessary steps to appropriately validate WPF models. The present study gives an extensive overview of the performance evaluation methods used for assessing the forecast accuracy of short-term statistical wind power forecast estimates, and the concept of robustness is introduced to determine the validity of a model over different wind power generation scenarios over the testing set. Finally, a numerical study using decomposition-based hybrid models is presented to analyse the robustness of the performance evaluation metrics under different conditions in the context of wind power forecasting. Data from Ireland are employed using two different resolutions to examine its influence on the forecast accuracy.

1. Introduction

Wind power forecasting (WPF) has established itself as one of the main challenges faced by the energy industry due to the stochastic character of the wind. As the penetration of wind power increases in the grid, accurate WPFs become more necessary since they lead to a greater performance of the energy market [1] and the operation of the grid [2]. Furthermore, WPFs errors cannot be prevented and consequently, they must be reduced and properly assessed to evaluate the validity of a WPF model.

WPF models can be broadly divided into physical and statistical methods. Physical forecasting models draw on meteorological information and specific site conditions at a current or future wind farm, combined with the laws of physics, to produce predictions. On the other hand, statistical models are built using historical data of the wind farm. Statistical methods such as time series modelling are used to predict future values of wind power output. Alternatively, statistical models are based on machine learning techniques, such as artificial neural networks (ANNs), or deep learning. Physical models do not require historical data from the wind farm.

Different aspects of WPF have been discussed in other papers. A chronological evolution of short-term WPF from a qualitative point of view is provided in [3]. [4] presents an overview of the main numerical wind prediction methodologies such as upscaling and downscaling and also WPF models based on statistical and machine learning methods. [5] introduces a collection of data-mining techniques to discover hidden patterns in a dataset. Some of these techniques are cluster analysis, to split a dataset into different groups with similar characteristics, or association analysis, to find relationships among observations. The performance of these techniques is evaluated for different time horizons. [6] analyses probabilistic methodologies to predict wind power generation and classifies them into three different categories: probabilistic forecasting [7], where the output is regarded as a random variable, risk index [8], where an index is used to define the level of uncertainty of the WPF, and scenario forecasting [9,10], where statistical scenarios are generated considering the spatial and temporal interdependence of prediction errors. [11] overviews combined forecasting techniques for wind speed and wind power prediction. Forecasting models are usually combined by estimating the output independently for each model and afterwards a weight coefficient depending on the efficiency of every model. [12] gives an in-depth analysis of wind power ramp forecasting, a subclass of wind power prediction that focuses on large and fast variations of wind power known as ramp events. [13] discusses the impact of different uncertainty sources of the wind power forecast, such as the weather conditions or the prediction algorithm.

* Correspondence to: University College Dublin, School of Mechanical and Materials Engineering, Engineering Building Belfield Dublin 4, Ireland.
E-mail address: vikram.pakrashi@ucd.ie (V. Pakrashi).

https://doi.org/10.1016/j.rser.2020.110515
Received 24 December 2019; Received in revised form 24 September 2020; Accepted 23 October 2020

Abbreviations

ACE	Average Coverage Error
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
CRPS	Continuous Ranked Probability Score
CWC	Coverage Width-based Criterion
EEMD	Ensemble Empirical Mode Decomposition
ELM	Extreme Learning Machine
FFNN	Feedforward Neural Network
IA	Index of Agreement
IS	Interval Sharpness
KDE	Kernel Density Estimation
LS-SVM	Least Squares Support-vector Machine
LUBE	Lower Upper Bound Estimation
MA	Moving Average
MAAPE	Mean Arctangent Absolute Percentage Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MDL	Minimum Description Length
MIMO	Multiple-Input Multiple-Output
NMAE	Normalized Mean Absolute Error
NRMSE	Normalized Root Mean Square Error
PDF	Probability Density Function
PI	Prediction Interval
PICP	PI Coverage Probability
PINAW	PI Normalized Average Width
PINC	PI Nominal Confidence
QR	Quantile Regression
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SC	Skill Score
SDE	Standard Deviation Error
SVM	Support-vector Machine
TSA	Time Series Analysis
VAR	Vector Autoregression
VMD	Variational Mode Decomposition
WPF	Wind Power Forecasting

None of the discussed studies goes into the limitations of evaluating WPFs, especially in terms of robustness, meaning that the forecast accuracy of a certain WPF model should not be affected by the wind power generation process, and therefore perform similarly for different scenarios of wind power generation [14]. Even though performance evaluation metrics are introduced in the literature to analyse the forecast accuracy of WPF models, the robustness of the models is often disregarded. The main contributions of this paper are the review of the main performance evaluation metrics used in the literature to assess WPF models and the empirical evaluation of this feature through a case study using a set of WPF models and two different multi-step ahead forecast strategies. Furthermore, forecasts are estimated using two different resolutions (10 and 60 min) to examine its influence in the evaluation of the models.

The remainder of this paper is organized as follows. Section 2 provides an overview of recently proposed WPF models. Section 3 presents the different performance evaluation methods for assessing WPFs. Section 4 presents different techniques used to estimate multi-step ahead forecasts that are considered for longer prediction horizons. Section 5 presents a numerical study with data from Ireland. Section 6 includes the concluding remarks of this paper.

2. Overview of WPF models

Traditionally, WPF forecasts have been provided as point or deterministic estimates, meaning that a single value is computed for every time step to forecast. Nonetheless, every prediction is associated with a certain degree of uncertainty which is impossible to reduce entirely. Probabilistic forecasts overcome this issue and allow to obtain a probabilistic estimate for future wind power outputs. Several representations for probabilistic estimates are found in the literature: predictive densities which represent the probability distribution of future outputs, quantiles that divide the probability distribution into intervals, and prediction intervals (PIs) which provide the range where a value will be located under a given distribution. The latter representation tends to be more appealing for end users, so usually PIs are provided to assess probabilistic estimates.

Comparing the different WPF models proposed in the literature is a challenging task as they are tested under different conditions. Firstly, the inputs given to the model differ, as univariate time series can be considered (only wind power data) or multivariate time series that reflect the dependency on other variables such as wind speed or wind direction. Other condition is the dataset in terms of its scale, as the model is fitted at either a turbine, farm or national production level, in terms of sample size, as it varies from a few months to approximately three years of data to benchmark the model, and in terms of time-resolution, usually from 10 min to 1 h, as intra-hour resolution data show higher volatility [15]. Forecasting competitions represent a good opportunity to compare different forecasting models [16,17], as they are evaluated under the same rules and using the same dataset.

Model performance is evaluated using deterministic and probabilistic estimates, or both. In this study, the aim is to create a benchmark of performance evaluation metrics from the existing literature. A brief overview of WPF models is presented to provide a context to the forecasting estimates. Fig. 1 shows an overview of the major categorizations of WPF models and estimates. Further details on the modelling of wind power are found in [18,19].

2.1. Time series analysis

In the realm of time-series analysis (TSA), autoregressive (AR) processes are modelled using a combination of previous variables, whereas moving average (MA) processes are modelled combining previous forecast errors. They can be merged together resulting in ARMA processes, and generalized to non-stationary processes by differencing the original time series, generating the known as ARIMA (autoregressive integrated moving average) models. Compared to ANNs and other machine learning models, TSA provides a well-established statistical framework that allow to draw conclusions more confidently [20]. In the field of WPF, vector autoregression (VAR) is a generalization of AR models that allows to predict WP for several wind farms considering their spatio-temporal dependencies. [21] tests a VAR-based method using 22 wind farms located in southeastern Australia, whereas [22] employs a similar method for 172 and 100 wind farms in France and Denmark respectively, proving to be effective for large datasets. ARFIMA (autoregressive fractionally integrated moving average model) models [23] are an extension of ARIMA that allows to characterize long-memory for time series. [24] proposes to use an ARFIMA process to model the linear component of WP time series, together with a least-squares support-vector machine (LS-SVM) to estimate the non-linear

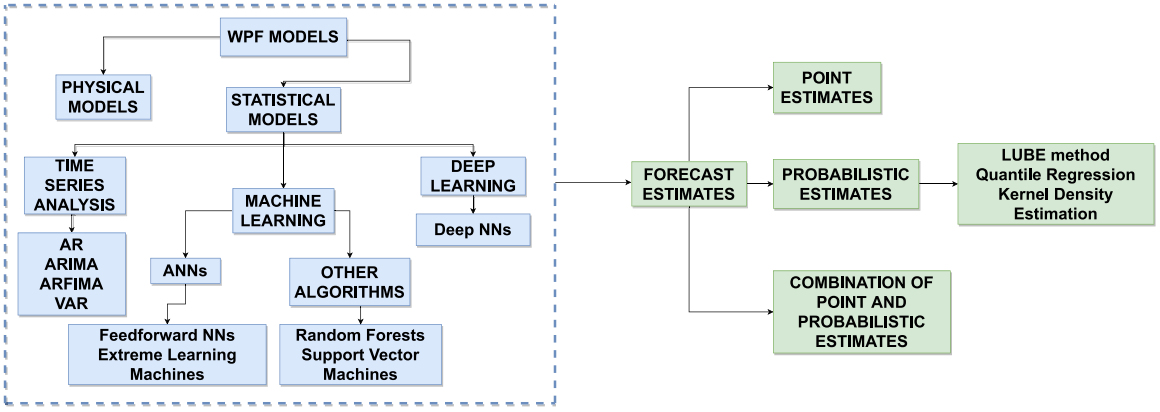

Fig. 1. Overview of WPF models and estimates.

component. [25] introduces a model where dynamic clustering and linear regression are its main features, and requiring less data in the training phase in comparison with ANNs or support-vector machines (SVMs) is its main advantage, being appropriate for new wind farms as the method can be applied with less than 1 month of data. [26] develops a non-parametric statistical model where meteorological changes are modelled by a Markov state transition process and the local dynamic behaviours by AR time series. [27] applies a hybrid method where an ARMA model is used as a first step to build a base model.

2.2. Machine learning

Most of the work on machine learning models is focused on ANNs. A basic feedforward neural network (FFNN) is defined as a classical network with a hidden layer between the input and output layers and uses the backpropagation training algorithm [28]. ANN methods used in the literature aim to improve forecast accuracy while reducing or maintaining the computational burden of the network by implementing improved optimization algorithms, using feature selection criteria to avoid the use of redundant data and avoid the propagation of errors generated by additional exogenous inputs, or using hybrid methodologies where the wind power time series is previously decomposed and the ANNs are applied to the resulting modes. For instance, improved optimization algorithms are found in [29], where the convergence and accuracy of the training algorithm is improved using an error feedback scheme, in [30] the clonal search algorithm contributes to capture nonlinearities in the data, and [31] uses a optimization algorithm where the concepts of evolutionary computing and particle swarm optimization are combined. Feature selections methods are implemented in [32], using genetic programming to prevent the propagation of errors of the predictors, in [33] data are clustered into groups of similar patterns and the best one is chosen and trained by an ANN, and in [34] a feature selection technique based on mutual information is used to select the most informative input variables to feed the ANN. Decomposition-based hybrid models decompose the WP time series before training the ANN and have shown a better forecasting performance in recent times [35]. Decomposition techniques such as wavelet transform are employed in [30,33], or variational mode decomposition (VMD) in [36]. Decomposition-based hybrid models are described in detail in Section 5.2.

ANNs can be also employed to built PIs. An interesting non-parametric approach is the method known as LUBE (lower upper bound estimation) method [37,38]. It consists of a FFNN with two outputs which represent the upper and lower boundaries of the constructed PI. The loss function is based on the two main properties of the PI: its coverage and its width. [39] adds a fuzzy-based loss function to the LUBE method to facilitate the adjustment of the NN parameters. The forecast accuracy of the method is demonstrated for two case studies where the PIs are evaluated changing the reference membership values and a set of NNs with structures from 5 to 15 neurons. [40] proposes a methodology inspired on the LUBE method in which the lower and upper boundaries of the PI are designed using a novel inter type-2 fuzzy model. Another alternative to build PIs is quantile regression (QR), a non-parametric method where the uncertainty is estimated by means of a set of forecast quantiles. QR is characterized by its free-distribution approach and its flexibility to include predictors [7]. Further details on QR are found in Section 5.2. This methodology can be combined easily with ANN models to extend deterministic estimates to PIs, such as [41], where QR is used to establish probabilistic estimates for a neural network based methodology. Additionally, a QR neural network is combined with another non-parametric approach known as kernel density estimation (KDE) in [42], where the predictions at different quantiles are used as an input for the KDE to model the wind power probability density function (PDF) information.

An alternative to the backpropagation algorithm to train faster ANNs are extreme learning machines (ELMs), a NN-based technique based on a single-hidden feedforward neural network (FFNN) where the weights between the input and the hidden layers are randomly assigned and never updated, accelerating the training rate of the network. ELM-based models found in the literature are [43], that proposes a bidirectional mechanism based on this technique, and [44] introduces a two-stage WPF model where the ELM is optimized by the grey wolf algorithm. Estimation of PIs using ELM can be found in [45–48].

Other WPF models using ANNs are combined with other techniques such as a SVM [49], a LS-SVM [50], or a Gaussian process [51]. Alternatively, other machine learning WFP models in the literature are [52], where and ensemble of decision trees and support vector regression is used, and [53], where deterministic estimates are generated by random forests and intervals by QR forests.

2.3. Deep learning

Deep learning based models are an extension of machine learning methods in the sense that networks are conformed by several layers that provide different interpretations of the data fed to the model. Details on the implementation of deep learning for energy forecasting can be found in [54]. [55] obtains WPFs with a deep learning based model using convolutional neural networks. [56] uses the LUBE method with a recurrent neural network (RNN) model instead of training the model with a FFNN. [57] proposes a deep neural network based ensemble technique combined with the concept of transfer learning to extend the knowledge gained training one wind farm to others.

3. Performance evaluation metrics of WP forecasts

The assessment of a forecasting model is a crucial step in its development to address its validity for estimating future values of wind power. The forecast accuracy of deterministic estimates is evaluated measuring the discrepancy between the forecast and actual values through several criteria. The evaluation of probabilistic estimates is a more challenging task, as the forecast cannot be compared directly to the actual values, and several properties of the forecast have to be addressed to verify the forecast accuracy of the model.

3.1. Accuracy of deterministic estimates

Many performance evaluation methods are used in the literature to assess the accuracy of deterministic estimates. The most common ones are shown down below.

- Mean absolute error (*MAE*):

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

- Root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{2}$$

- Mean absolute percentage error (*MAPE*):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\% \tag{3}$$

- Standard deviation error (*SDE*):

$$SDE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\epsilon_i - \bar{\epsilon})^2} \tag{4}$$

- Bias:

$$BIAS = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{y}_i \tag{5}$$

- Index of Agreement (*IA*):

$$IA = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \tag{6}$$

where N is the number of samples, y_i is the actual value, $\hat{y}_i$ the predicted value, $\bar{y}$ is the mean value of the real values, $\epsilon_i = y_i - \hat{y}_i$ is the prediction error (also known as residual), and $\bar{\epsilon}$ is the average value of the errors. Table 1 summarizes the performance evaluation metrics used in the literature for deterministic wind power estimates.

Eq. (1) shows the mean absolute error (*MAE*). It is defined as the average value of the predictions errors in absolute values. The root mean square error (Eq. (2)) depicts the standard deviation of the residuals. Normalized versions of the *MAE* (*NMAE*) and the *RMSE* (*NRMSE*) are commonly used in the literature as well. While both *MAE* and *RMSE* are suitable indicators for assessing the performance of a model, the *RMSE* should be preferred when the model errors follow a Gaussian distribution [64].

Another statistical measure is the *MAPE* (Eq. (3)). It quantifies the accuracy as a percentage of the error. However, the *MAPE* produces very large values when the actual values are close to zero and is undefined when the actual value is equal to zero. Alternative versions of the *MAPE* have been proposed to prevent this shortcoming. For instance, the mean absolute scaled error (*MASE*) is an alternative defined as the *MAE* of the forecast values scaled by the *MAE* of the in-sample naïve forecast [65]. It is specially useful for wind power time series, as there are periods where a wind farm does not generate any power. The mean arctangent absolute percentage error (*MAAPE*) is another alternative option to the *MAPE* [66]. It transforms the *MAPE* using the arctangent function. Its main advantage is the preservation of the characteristics of the *MAPE* while overcoming the limitations of the *MAPE*.

The prediction error can be decomposed into the random error, which is inherently unpredictable, and the systematic error, which occurs due to inaccuracies in the system. The standard deviation of errors (Eq. (4)) addresses the random component of the prediction error, whereas the bias (Eq. (5)) deals with the systematic component [67].

Lastly, another common metric in the literature to assess deterministic estimates is the index of agreement (Eq. (6)). It was originally proposed in [68] and refined versions of this index have been developed since then [69]. It measures to which degree the predictions are error-free and takes values between zero when the adjustment between predictions and observations is null, and one when the predictions fully pair with the actual values.

Another alternative to examine model performance is to use probabilistic statistical measures that address not only the performance but also the complexity of the model. Some of them are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Minimum Description Length (MDL).

In order to facilitate the comparison with benchmark models, the improvement of a technique is defined by means of the relation [67]:

$$\text{Improvement} = 1 - \frac{M}{M_{ref}} \tag{7}$$

where M is the value of the selected measure for a specific model, and M_{ref} is the value of the same measure for the benchmark model. Comparing different models become an issue as there is not a unified criterion for selecting benchmarks to evaluate them. Typically, models are compared to the persistence model, as it is a requirement to outperform it to be considered skillful, and state-of-the-art methods such as neural networks.

3.2. Accuracy of probabilistic estimates

A PI is an interval that gives the expectation of where a future value will fall with a specified probability. Therefore, the PI relies on the significance level α. The probability that a future wind power output y_i lies within the PI is known as Prediction interval nominal confidence (*PINC*):

$$PINC = 100(1-\alpha)\% \tag{8}$$

Taking this into consideration, a PI for a future time step i and a significance level α is defined as:

$$\hat{I}_i^{\alpha} = \hat{U}_i^{\alpha} - \hat{L}_i^{\alpha} \tag{9}$$

where $\hat{U}_i^{\alpha}$ and $\hat{L}_i^{\alpha}$ are the upper and lower boundaries of the PI respectively. The most common metrics defined for probabilistic estimates are shown down below.

- Prediction interval coverage probability (*PICP*):

$$PICP = \frac{1}{N}\sum_{i=1}^{N} c_i \tag{10}$$

where N is the number of samples and c_i is

$$c_i = \begin{cases} 1, & \text{if } y_i \in \hat{I}_i^{\alpha} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

- Prediction interval normalized average width (*PINAW*):

$$PINAW = \frac{1}{NR}\sum_{i=1}^{N}\hat{I}_i^{\alpha} \tag{12}$$

where R is the range of the target variable.

Table 1
Performance evaluation metrics for deterministic estimates.

Reference	Year	MAE	NMAE	RMSE	NRMSE	MAPE	SDE	Bias	IA	Others
Dowell et al [21]	2015	×		×						
Messner et al [22]	2019	×		×				×		
Yuan et al [24]	2017	×		×		×				
Ozkan et al [25]	2015		×		×					
Jiang et al [27]	2017	×	×	×						
Chang et al [29]	2017					×				×
Chitsaz et al [30]	2015		×		×					
Osório et al [31]	2015		×		×	×				
Zameer et al [32]	2017	×		×			×			
Azimi et al [33]	2016	×		×					×	
Li et al [34]	2015				×	×				
Naik et al [36]	2018	×		×		×				
Haque et al [41]	2014		×		×	×				
He et al [42]	2018			×		×				×
Zhao et al [43]	2016		×		×		×	×		×
Hao et al [44]	2019	×		×		×			×	
Buhan et al [49]	2015		×							
Liu et al [50]	2017		×		×	×				
Lee et al [51]	2013	×		×		×	×	×		×
Heinermann et al [52]	2016									×
Lahouar et al [53]	2017	×	×	×		×				×
Qureshi et al [57]	2017	×		×			×			
Y. Zhang et al [58]	2016			×						
Yan et al [59]	2016		×		×					
Yang et al [60]	2015	×		×						
Y. Wang et al [61]	2017		×		×					×
Han et al [62]	2015	×		×						
Zjavka et al [63]	2018	×		×						×

Table 2
Performance evaluation metrics for probabilistic estimates.

Reference	Year	PICP	PINAW	CWC	ACE	IS	CRPS	SC	Others
Dowell et al [21]	2015						×		×
Xie et al [26]	2018	×						×	
Khosravi et al [37]	2013	×	×	×					
Quan et al [38]	2013	×	×	×					
Kavousi-Fard et al [39]	2015	×	×						
Zou et al [40]	2019	×	×	×					×
Haque et al [41]	2014							×	×
He et al [42]	2018	×	×						
G. Zhang et al [45]	2014	×	×						
Wan et al [46]	2016							×	×
Mahmoud et al [47]	2018	×	×			×			×
Afshari-Igder et al [48]	2018	×				×			
Lahouar et al [53]	2017								×
H. Wang et al [55]	2017				×	×	×		
Shi et al [56]	2017	×	×	×					×
Y. Zhang et al [58]	2016								×
Yang et al [60]	2015						×		
Y. Wang et al [61]	2017	×	×	×	×	×			
Gallego-Castillo et al [70]	2016						×		
Lin et al [71]	2018						×		×
Khorramdel et al [72]	2018	×	×			×			
Alessandrini et al [73]	2015						×		×

- Coverage width-based criterion (*CWC*):

$$CWC = PINAW\left[1 + \gamma(PICP)e^{-\eta(PICP-\mu)}\right] \tag{13}$$

where $\gamma(PICP)$ is a step function dependent on the values of PICP and μ:

$$\gamma(PICP) = \begin{cases} 0, & \text{if } PICP \geq \mu \\ 1, & \text{if } PICP < \mu \end{cases} \tag{14}$$

- Average coverage error (*ACE*):

$$ACE = PICP - PINC \tag{15}$$

- Interval sharpness (*IS*):

$$IS = \frac{1}{N}\sum_{i=1}^{N} b_i \tag{16}$$

where b_i is

$$b_i = \begin{cases} -2\alpha\hat{I}_i^{\alpha} - 4(\hat{L}_i^{\alpha} - y_i), & \text{if } y_i < \hat{L}_i^{\alpha} \\ -2\alpha\hat{I}_i^{\alpha}, & \text{if } y_i \in \hat{I}_i^{\alpha} \\ -2\alpha I_i^{\alpha} - 4(y_i - \hat{U}_i^{\alpha}), & \text{if } y_i > \hat{U}_i^{\alpha} \end{cases} \tag{17}$$

- Continuous ranked probability score (*CRPS*):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_0^{P_{max}} [CDF_i - H(y - y_i)]^2 dy \tag{18}$$

where CDF_i is the cumulative form of the distribution and H is the Heaviside step function:

$$H = \begin{cases} 0, & \text{if } y < y_i \\ 1, & \text{otherwise} \end{cases} \tag{19}$$

- Skill Score (*SC*):

$$SC = \frac{1}{N}\sum_{i=1}^{N} SC_i = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{M}(\xi^{\alpha_j} - \alpha_j)(y_i - q_i^{\alpha_j})\right] \tag{20}$$

where SC_i is a set of quantiles on a single time step i, α_j is the quantile proportion, $q_i^{\alpha_j}$ is the quantile forecast, and ξ^{α_j} is an indicator variable denoted by

$$\xi^{\alpha_j} = \begin{cases} 1, & \text{if } y_i < q_i^{\alpha_j} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

Table 2 shows the application of these performance evaluation metrics for probabilistic estimates in the recent literature.

The two main features of a PI, the most common representation for probabilistic estimates, are its reliability and its informativeness: a PI will be reliable when the actual wind power output falls within the interval, whereas it will be informative depending on its width. Ideally, the PI should be as narrow as possible to facilitate the decision-making process. The uncertainty of the prediction can be attributed to several sources. One source is the model uncertainty (or epistemic uncertainty), which occurs due to a misspecification of the forecasting model or the parameters of the model. The other main source is the data uncertainty (or aleatoric uncertainty), which quantifies the inherent noise of the observations.

The *PICP* (Eq. (10)) is a metric that measures exclusively the reliability of the PI. It accounts for the average of target values covered by the interval. Its counterpart in terms of width is the *PINAW* (Eq. (12)). The *PICP* and the *PINAW* can be further merged with the *CWC* (Eq. (13)). In this equation, η and γ are two controlling hyperparameters that determine how much invalid PIs are penalized. Alternative versions of the *CWC* have been introduced in the literature. For instance, the new *CWC* is proposed in [56], which includes a new term designed to take better into consideration the information provided by the actual measurements. Another alternative is presented in [45], which considers an additional function to account for those samples that lie beyond the interval.

Another parameter that describes the reliability of a PI is the *ACE* (Eq. (15)). It is defined as the deviation between the *PICP* and the *PINC*. Smaller deviations indicate more reliable PIs. This metric provides additional information compared to the *PICP* since a larger *PICP* is not necessarily better for a given *PINC* [74].

The interval sharpness (also known as Winkler score) (Eq. (16)) evaluates the PI in terms of its width [75]. Narrower intervals are rewarded by this metric, whereas those PIs where the observations do not lie inside are penalized.

The *CRPS* (Eq. (18)) is a global criterion as it assesses both features simultaneously. This metric is equivalent to the *MAE* when a forecast generates a deterministic estimate. It differs from other metrics as the *CRPS* assesses cumulative distribution functions. Lower scores of the *CRPS* mean a better performance of the model.

The Eq. (20) denotes the scoring rule proposed by [76] when probabilistic forecasts are estimated by non-parametric models and are represented by a set of quantile forecasts. Its orientation is positive and a value of zero represents a perfect forecast.

Comparison with benchmark models for probabilistic estimates can be performed by using the same relation presented for deterministic estimates in Eq. (7).

4. Prediction horizon for WPF estimates

The prediction horizon is one of the main aspects to consider when a WPF model is developed. Very-short term forecasts consider predictions up to 30 min ahead. The main applications of these forecasts are wind farm control and operation of reserves [77]. Short-term forecasts take into account predictions from hours to a few days and are an indispensable tool for power system management and energy trading. The use of additional exogenous inputs such as meteorological data may be considered to train a statistical model for this horizon as the dynamics of wind power generation become significant and could potentially lead to a better performance of the model. Nonetheless, the use of these data could increase the computational complexity of the model [78] and these datasets are associated with their own prediction error that will affect the prediction accuracy [79]. Alternatively, physical models are used for short-term forecasting as well in the absence of an actual wind farm, although they are more computationally expensive compared to statistical models. Longer-term forecasts usually make use of physical methods and are used to make decisions on unit commitment or maintenance scheduling [80].

For longer prediction horizons from very short-term and short-term statistical models, it is necessary to consider multi-step ahead forecasts. A multi-step ahead forecast estimates the next H steps $[y_{t+1}, \dots, y_{t+H}]$ of a time series. There are several strategies to approach this matter [81, 82]: the recursive, the MIMO (Multiple-input Multiple-output) and the direct strategies are the most commonly used approaches to estimate multi-step ahead predictions.

In the recursive strategy [81], also known as iterated or multi-state strategy, the model is trained to compute one-step ahead forecasts. Afterwards, the next steps are predicted iteratively using the previous one-step ahead forecasts as inputs. This strategy is sensitive to larger prediction horizons, as the errors of the predictions accumulate for every iteration.

$$\hat{y}_{t+h} = \begin{cases} f(y_t, \dots, y_{t-d+1}) & \text{if } h = 1, \\ f(\hat{y}_{t+h-1}, \dots, \hat{y}_{t+1}, y_t, \dots, y_{t-d+h}) & \text{if } h \in \{2, \dots, d\} \\ f(\hat{y}_{t+h-1}, \dots, \hat{y}_{t+h-d}) & \text{if } h \in \{d+1, \dots, H\} \end{cases} \tag{22}$$

where d denotes the number of steps used in the input set.

The MIMO strategy [83] produces a vector with the whole sequence of outputs training a single model.

$$[\hat{y}_{t+H}, \dots, \hat{y}_{t+1}] = f(y_t, \dots, y_{t-d+1}), \tag{23}$$

The direct strategy [81] consists of training H different models independently, one for each horizon. While it prevents the accumulation of errors, it neglects the dependencies between the H forecasts and it carries a larger computational cost to train every model separately.

$$\hat{y}_{t+h} = f_h(y_t, \dots, y_{t-d+1}), \qquad h \in \{1, \dots, H\} \tag{24}$$

Further combinations of these models lead to other strategies. For instance, the DirRec strategy [84] combines elements from the recursive and direct approaches, and the DIRMO strategy [85] presents a trade-off between the direct and the MIMO strategies.

5. Numerical study

In this section, the features of the performance evaluation metrics for WPF models have been investigated considering data from a single wind-farm, modelled using a set of decomposition-based hybrid models. Firstly, the dataset and the forecasting models are described, followed by a discussion of the features of the metrics over the testing set.

5.1. Dataset

The dataset used for the case study contains measurements from a wind farm located in Ireland. Measurements are collected every 10 min between January 2017 and December 2017 (Fig. 2). As shown in this figure, wind power generation shows a large variability as it is influenced by wind and other meteorological variables, as well as human activities such as maintenance operations. Less than 1% of the values are missing and have been reconstructed considering the previous and posterior values. Simulations are run for one wind turbine with the reconstructed dataset with a temporal 10-min resolution and a resampled dataset with a 1-h resolution. The dataset has been divided

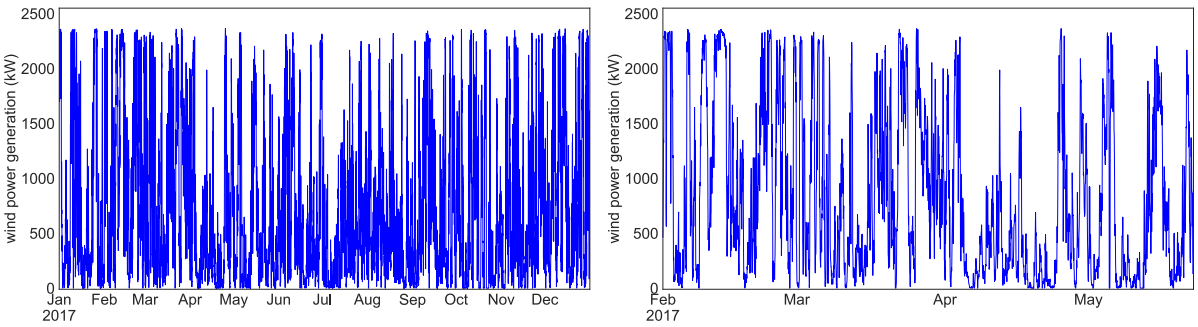

Fig. 2. Historical wind power generation during the year 2017 (left) and a sample of the wind power generation time series from February to May (right). Data are shown with a temporal resolution of 60 min.

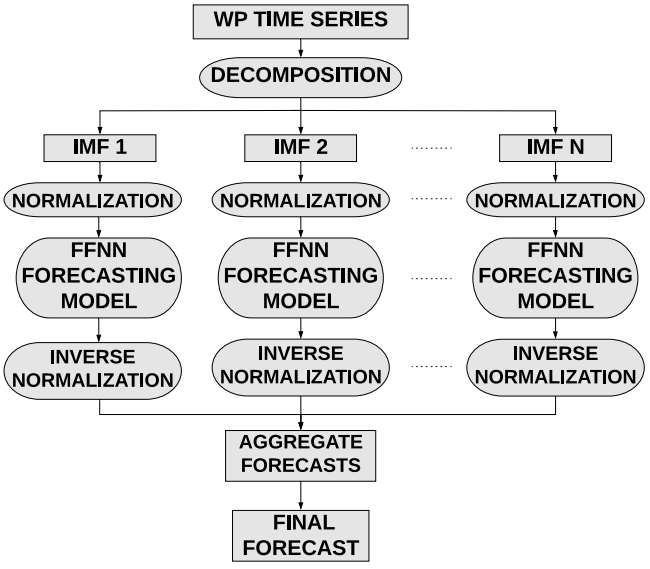

Fig. 3. Flowchart of the WP forecasting model.

Table 3
Sample size of training, validation and testing sets and summary statistics of the datasets.

	Low-res set	High-res set
Resolution	60-min	10-min
Train	7296	43776
Validation	1080	6480
Test	384	2304
Mean (kW)	726.49	726.49
Std (kW)	696.15	715.30
Min (kW)	0	0
Q1 (kW)	157.13	140
Q2 (kW)	478	462
Q3 (kW)	1134	1151
Max (kW)	2364.3	2365

into training, validation and testing sets to train the model and test the accuracy of the forecasts provided by the models. Table 3 shows the number of samples for each set considering the low-resolution (1-h) and high-resolution (10-min) data and the summary statistics for both sets of data. The location of the wind farm is not disclosed due to confidentiality reasons.

5.2. Modelling methodology

The performance evaluation metrics existing in the literature show that decomposition-based hybrid models contribute to a better forecasting accuracy for WPF [35]. Two different techniques are applied to decompose the wind power time series: ensemble empirical mode decomposition (EEMD) and VMD. Additionally, two multi-step forecast strategies (MIMO and Recursive) are implemented. In total, six models are employed (Table 4), including two FFNN models where the WP time series is not decomposed.

Table 4
WPF models used for the numerical study.

Decomposition technique	Training model	Multi-step strategy
VMD	FFNN	MIMO
VMD	FFNN	Recursive
EEMD	FFNN	MIMO
EEMD	FFNN	Recursive
–	FFNN	MIMO
–	FFNN	Recursive

EEMD [86] is a non-linear signal processing technique in which the time series is decomposed into a set of stationary modes and mitigate mode mixing issues existing in the standard EMD approach [87]. EEMD obtains the modes, known as intrinsic mode functions (IMFs), as the mean value of the modes generated by an ensemble of various noise-added copies of the original signal.

VMD [88] is another method to decompose a signal into its principal modes. The modes are estimated concurrently by looking for a set of modes and their centre frequencies all together to reproduce the signal. The bandwidth of each mode is estimated by a constrained variational optimization process: first, the Hilbert transform is employed to determine the associated analytical signal and consequently a unilateral frequency spectrum; second, the mode's frequency spectrum is shifted to baseband by mixing with an exponential tuned to the respective estimated centre frequency; and third, the H^1 Gaussian smoothness of the demodulated signal is used to identify the bandwidth of the mode. This process is transformed into an unconstrained problem by introducing a penalty term and Lagrangian multipliers λ as it follows:

$$L(\{u_k\},\{\omega_k\},\lambda) = \alpha \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| y(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 + \left\langle \lambda(t), y(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \quad (25)$$

where $y(t)$ is the original time series, $\{u_k\}$ the set of all modes, $\{\omega_k\}$ the set of the respective centre frequencies, $\delta(t)$ the Dirac function, and α the balancing parameter of the data fidelity constraint. Finally, this equation is solved using the alternate direction method of multipliers [89].

Using either EEMD or VMD, every resulting mode is trained using an FFNN and afterwards, the signal is reconstructed by aggregating all forecasts (Fig. 3). The FFNN is defined as a classical network with a hidden layer between the input and output layers and uses the backpropagation training algorithm [28], in which a Rectified Linear Unit (ReLU) is used as activation function in the network. In order

Table 5
Results for deterministic estimates.

Forecast horizon: 6-h ahead							
Low-resolution dataset (1-h resolution)							
Method	Strategy	NMAE (%)	NRMSE (%)	MAPE (%)	NBias (%)	NSDE (%)	IA
EEMD-FFNN	MIMO	10.40	14.33	29.80	4.12	13.72	0.936
EEMD-FFNN	Recursive	10.80	14.66	31.45	−2.09	14.51	0.935
VMD-FFNN	MIMO	**1.77**	**2.44**	**12.05**	**−0.58**	**2.37**	**0.998**
VMD-FFNN	Recursive	4.51	5.45	20.25	4.38	3.23	0.992
FFNN	MIMO	21.08	28.41	44.91	5.98	27.97	0.706
FFNN	Recursive	21.87	30.12	48.61	2.35	30.02	0.730
High-resolution dataset (10-min resolution)							
Method	Strategy	NMAE (%)	NRMSE (%)	MAPE (%)	NBias (%)	NSDE (%)	IA
EEMD-FFNN	MIMO	11.16	15.50	28.69	**1.03**	15.46	0.929
EEMD-FFNN	Recursive	24.91	33.02	56.97	23.51	23.19	0.675
VMD-FFNN	MIMO	**8.27**	**10.77**	**25.19**	−3.61	**10.14**	**0.969**
VMD-FFNN	Recursive	13.65	19.02	35.50	9.49	16.48	0.890
FFNN	MIMO	22.17	30.55	46.84	7.48	29.62	0.689
FFNN	Recursive	34.26	46.11	74.19	34.05	31.10	0.478
Forecast horizon: 24-h ahead							
Low-resolution dataset (1-h resolution)							
Method	Strategy	NMAE (%)	NRMSE (%)	MAPE (%)	NBias (%)	NSDE (%)	IA
EEMD-FFNN	MIMO	18.18	22.33	35.84	**1.54**	22.27	0.765
EEMD-FFNN	Recursive	24.69	33.03	55.89	20.71	25.73	0.582
VMD-FFNN	MIMO	**5.95**	**7.57**	**18.99**	1.66	7.38	**0.982**
VMD-FFNN	Recursive	8.81	10.48	25.78	−7.68	**7.13**	0.969
FFNN	MIMO	26.66	34.93	49.68	12.66	32.56	0.379
FFNN	Recursive	28.74	36.92	58.63	12.50	34.74	0.434
High-resolution dataset (10-min resolution)							
Method	Strategy	NMAE (%)	NRMSE (%)	MAPE (%)	NBias (%)	NSDE (%)	IA
EEMD-FFNN	MIMO	**20.28**	**25.43**	38.52	**1.06**	**25.32**	0.689
EEMD-FFNN	Recursive	43.49	53.73	**36.47**	−39.05	36.91	0.522
VMD-FFNN	MIMO	20.39	25.94	40.41	4.90	25.48	**0.713**
VMD-FFNN	Recursive	29.40	35.77	38.97	−11.20	33.97	0.589
FFNN	MIMO	28.62	38.25	56.42	17.63	33.95	0.389
FFNN	Recursive	63.79	71.43	31.31	−63.79	32.14	0.411

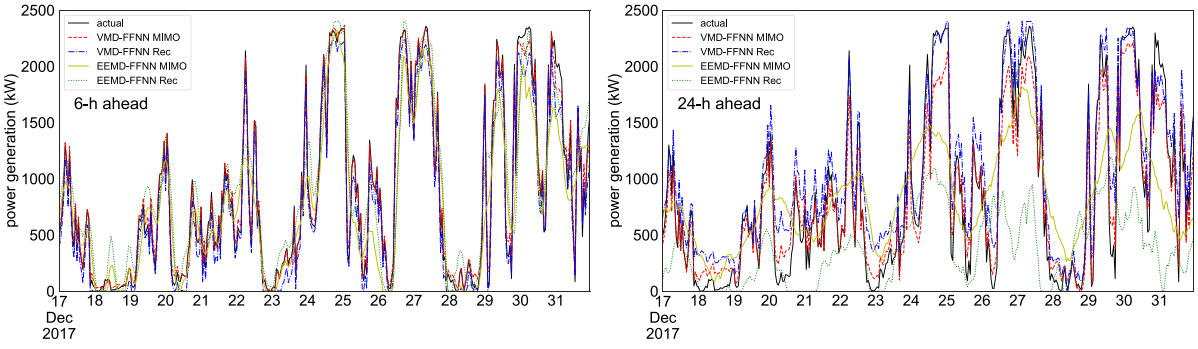

Fig. 4. Performance of the selected models for 6-h and 24-h ahead deterministic estimates. Data are shown with a temporal (60-min) resolution.

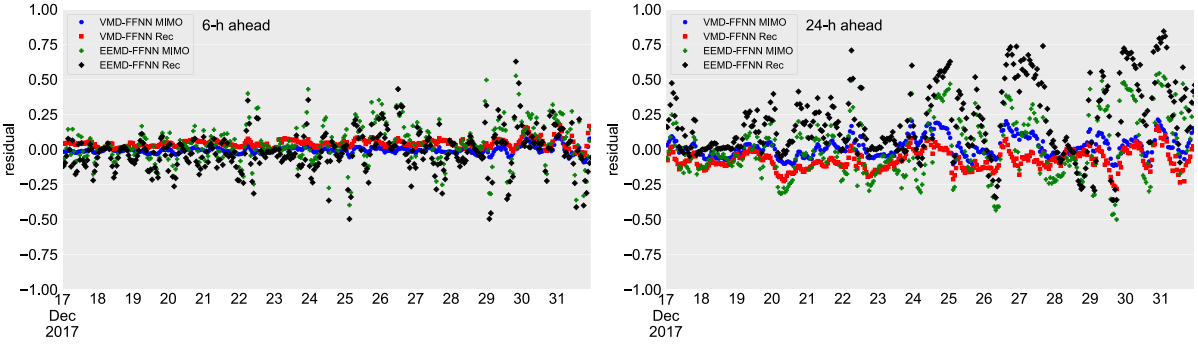

Fig. 5. Residuals for 6-h and 24-h ahead deterministic estimates. Data are shown with a temporal (60-min) resolution.

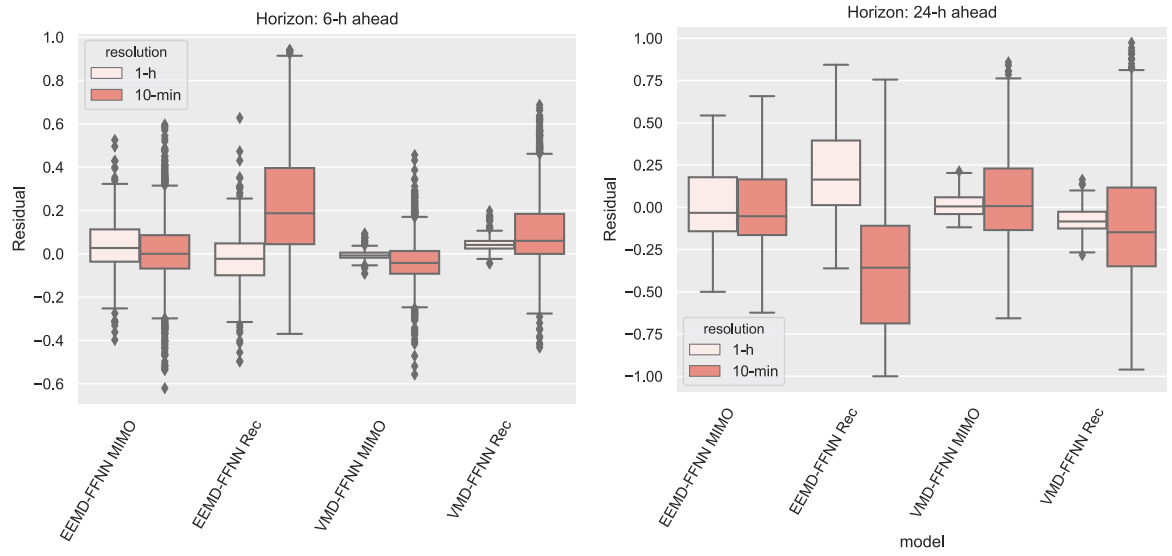

Fig. 6. Distribution of the residuals for 6-h and 24-h ahead predictions.

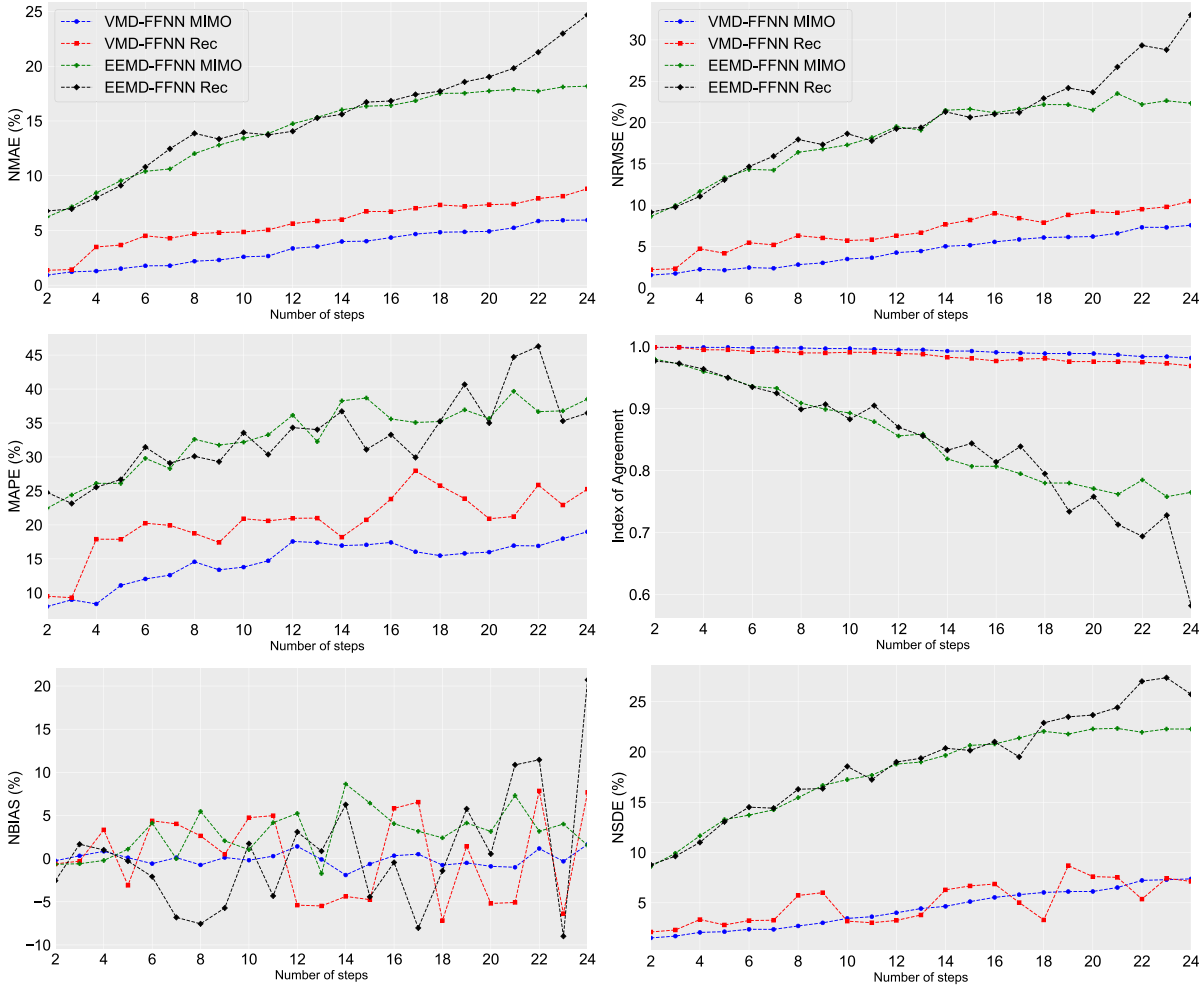

Fig. 7. Performance evaluation metrics for deterministic estimates with respect to the number of steps ahead (low resolution dataset).

to facilitate the training of the network, data from every mode are normalized before the training using the following expression:

$$y_{norm} = \frac{y - y_{min}}{y_{max} - y_{min}} \tag{26}$$

where y is a data point of a given mode, y_{max} is the maximum value, and y_{min} is the minimum value of that mode. As the outputs are normalized as well, the scaling is undone (inverse normalization) before aggregating the forecast outputs of each mode. In every case for this

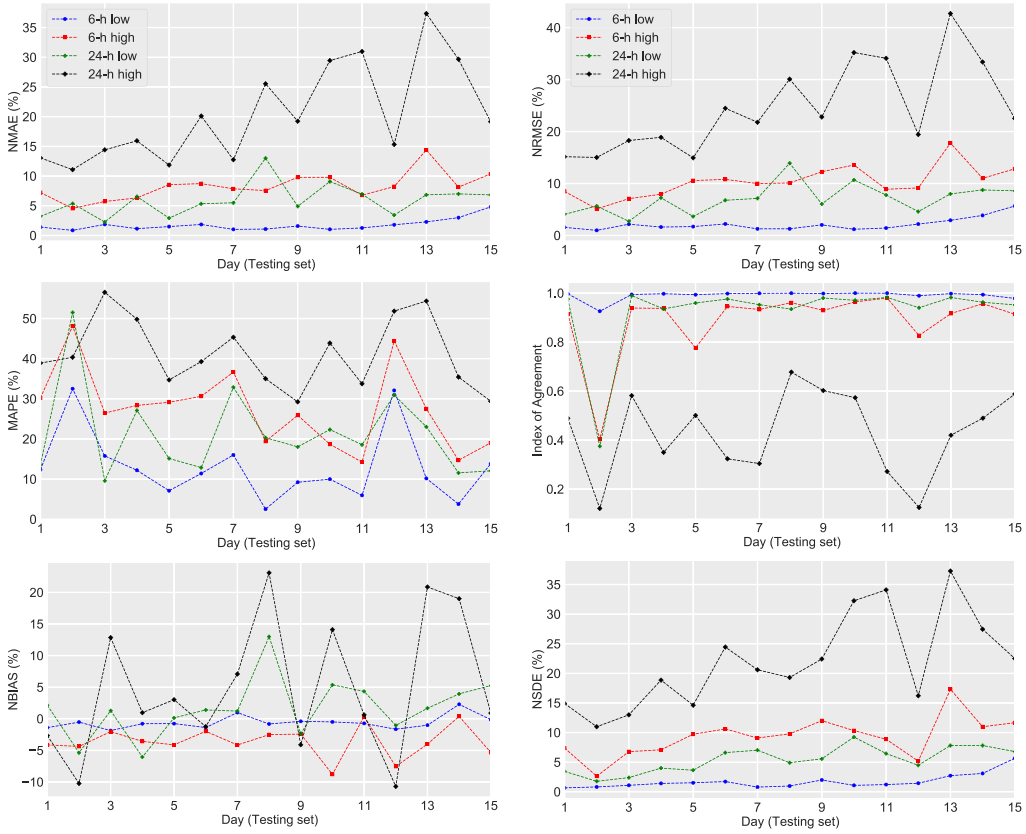

Fig. 8. Daily performance of metrics for deterministic estimates (VMD-FFNN MIMO model).

study, the vector input has the same size as the vector output, which will be as long as the prediction horizon.

Lastly, the boundaries of the PIs are estimated by quantile regression, using an asymmetric loss function (also known as pinball loss function) that depends on the required quantile τ:

$$\rho_\tau(\epsilon) = \begin{cases} \tau\epsilon, & \text{if } \epsilon \geqslant 0 \\ (\tau - 1)\epsilon, & \text{otherwise} \end{cases} \tag{27}$$

and taking it into consideration, the error function to be minimized is

$$E_\tau = \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(y(i) - \hat{y}_\tau(i)) \tag{28}$$

where $y(i)$ is the target value at a time i and $\hat{y}_\tau(i)$ is the conditional τ-quantile at the same time. By doing so, the conditional quantiles are estimated instead of the conditional mean, allowing to compute PIs. Further information can be found in [90,91].

Deterministic estimates are obtained using only one output for every FFNN, whereas the boundaries of PIs are estimated by QR as described in Eqs. (27) and (28). This technique has been chosen for the case study to provide interval estimates as it is a well-established technique in the field of WPF [7,92]. Further methodologies based on QR can be found in the literature [41,46]. The final prediction will be a single output for every step for deterministic estimates and the lower and upper boundaries of the interval in the case of PIs.

5.3. Results

The decomposition-based hybrid models previously described are used to obtain forecast estimates, that will allow to produce numerical values for the performance evaluation metrics. In this study, the metrics not only evaluate the general forecast accuracy of the models, as usually considered during the model development stage, but also in terms of time-resolution, robustness, and prediction horizon length. In the case of probabilistic estimates, PIs are chosen to present the uncertainty of the forecast as they are the most widespread representation. Hence the CRPS and SC are not considered in the study, as they are used to assess other representations of probabilistic estimates.

5.3.1. Deterministic predictions

Forecasts are estimated 6-h and 24-h ahead. These horizons are important for activities such as energy trading, where an initial forecast is usually provided 24-h ahead and subsequently corrected between 6- and 8-h ahead. Table 5 shows the results of the performance evaluation measurements for 6-h and 24-h ahead forecasts.

The values of *MAE*, *RMSE*, *BIAS*, and *SDE* are normalized by the capacity of the wind turbine to facilitate their understanding and the assessment of the model errors. The values of *MAPE* and *IA* are not normalized, as *MAPE* is established by definition as a percentage, and *IA* only takes values between zero and one. The lower scores for *NMAE* and *NRMSE* indicate that overall the VMD-FFNN MIMO model produces better forecasts considering every source of error. Additionally, the better scores for the *NBIAS* and *NSDE* indicate that this model deals better with the systematic and random error separately. The numerical values of the metrics are larger for 24-h ahead forecasts, indicating that the forecast accuracy is lower. Values close to one for the *IA*, such as the VMD-FFNN MIMO models for both 6-h and 24-h ahead forecasts using the low-resolution dataset, indicate that the forecasts have a low degree of error. Compared to the rest of the metrics, *MAPE* shows unsteady values that are not consistent with the rest of metrics evaluating the overall forecast accuracy (*NMAE*, *NRMSE*, and *IA*). As observed in Fig. 4, the 6-h ahead forecasts are more accurate than 24-h ahead forecasts, as the forecast accuracy is lower for larger prediction horizons. Fig. 5 shows the residuals in the 1-h resolution testing set. In both scenarios, the prediction error shows less variability when the VMD technique is applied to decompose the wind power time series.

Table 6
Results for 6-h ahead prediction intervals.

Confidence level: 99%: Low-resolution dataset (1-h resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	**100**	99.66	0.997	**0.01**	−47.84
EEMD-FFNN	Recursive	95	76.77	6.440	−0.04	−78.45
VMD-FFNN	MIMO	**100**	77.95	0.780	**0.01**	−37.42
VMD-FFNN	Recursive	**100**	**60.05**	**0.601**	**0.01**	**−28.82**
FFNN	MIMO	93.61	92.68	14.641	−0.05	−55.95
FFNN	Recursive	63.89	80.69	3.3e7	−0.35	−394.49
Confidence level: 99%: High-resolution dataset (10-min resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	**100**	96.88	0.969	**0.01**	−46.50
EEMD-FFNN	Recursive	96.25	81.20	1.881	−0.03	−80.02
VMD-FFNN	MIMO	**100**	90.32	**0.903**	**0.01**	**−43.35**
VMD-FFNN	Recursive	88.98	**40.14**	1.495	−0.1	−92.97
FFNN	MIMO	99.8	98.77	0.988	0.01	−47.51
FFNN	Recursive	41.48	67.26	2.1e12	−0.57	−1096.71
Confidence level: 95%: Low-resolution dataset (1-h resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	**100**	81.96	0.820	**0.05**	−196.70
EEMD-FFNN	Recursive	88.61	41.29	10.486	−0.06	−184.08
VMD-FFNN	MIMO	**100**	39.75	0.397	**0.05**	−95.39
VMD-FFNN	Recursive	**100**	**31.09**	**0.311**	**0.05**	**−74.61**
FFNN	MIMO	93.06	83.43	3.04	−0.01	−263.53
FFNN	Recursive	43.89	50.51	6.3e10	−0.5	−723.95
Confidence level: 95%: High-resolution dataset (10-min resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	**100**	84.76	0.848	**0.05**	−203.42
EEMD-FFNN	Recursive	45.74	32.07	44.521	−0.49	−850.53
VMD-FFNN	MIMO	**100**	68.49	**0.685**	**0.05**	**−164.37**
VMD-FFNN	Recursive	57.78	**26.06**	11.038	−0.37	−346.90
FFNN	MIMO	90.97	85.55	7.265	−0.04	−269.465
FFNN	Recursive	27.78	44.97	1.8e14	−0.67	−1373.64
Confidence level: 80%: Low-resolution dataset (1-h resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	96.67	51.59	0.516	**0.17**	−528.03
EEMD-FFNN	Recursive	56.11	18.44	28395.9	−0.24	−526.32
VMD-FFNN	MIMO	**100**	21.34	0.213	0.2	−204.90
VMD-FFNN	Recursive	96.67	**14.26**	**0.143**	**0.17**	**−146.72**
FFNN	MIMO	71.94	56.99	32.57	−0.08	−854.34
FFNN	Recursive	33.61	32.82	3.9e9	−0.46	−1234.63
Confidence level: 80%: High-resolution dataset (10-min resolution)						
Method	Strategy	PICP (%)	PINAW (%)	CWC	ACE	IS
EEMD-FFNN	MIMO	99.44	58.90	0.589	**0.19**	−567.42
EEMD-FFNN	Recursive	15.88	15.94	81.20	−0.64	−1477.55
VMD-FFNN	MIMO	**99.95**	47.93	**0.479**	**0.19**	**−460.54**
VMD-FFNN	Recursive	36.76	**11.50**	8.798	−0.43	−511.88
FFNN	MIMO	77.18	59.92	3.059	−0.03	−905.04
FFNN	Recursive	21.67	16.84	7.8e11	−0.58	−1751.05

Spikes are visible when using EEMD, as the prediction error is larger due to the less ability of EEMD to predict sudden changes in wind power correctly.

In terms of the effect of time-resolution on the forecasts, the larger volatility existing in higher resolution wind power data reduces the forecast accuracy of WPF models. Fig. 6 depicts the distribution of the residuals for every scenario. The larger spread observed for the high-resolution data comes mostly from the volatility of these data, as their intrinsic characteristic are harder to capture for the model, and it is affected as well by the number of steps ahead to predict as the errors accumulate for every step. Low-resolution residuals show a normal distribution in most of the cases, except for the EEMD models for 24-h ahead, that are slightly skewed to the right. Furthermore, the 6-h ahead predictions present a few outliers, as most of the residuals are centred around the median. The prediction errors from the high-resolution dataset have different patterns. A right-skewed distribution is observed in the EEMD-FFNN Recursive and the VMD-FFNN Recursive models for 6-h ahead forecasts, and in the EEMD-FFNN MIMO, VMD-FFNN MIMO and VMD-FFNN Recursive models for 24-h ahead forecasts. A normal distribution is visible for the EEMD-FFNN MIMO and VMD-FFNN MIMO models, although there are a considerable amount of outliers at the end of both tails. Only the EEMD-FFNN Recursive model shows a skewness to the left. As the error distribution seems to be influenced by the model and the resolution, *RMSE* can be in some scenarios a more suitable metric compared to *MAE* to examine the accuracy of deterministic estimates of wind power forecasts [64], since it provides not only information about the performance of the model but the error distribution. Additionally, as the *RMSE* is related to the second moment of the error, it represents more accurately the presence of larger residuals.

The evolution of the performance evaluation metrics with respect to the number of steps ahead predicted for the low-resolution dataset is shown in Fig. 7. As every step represents a 60-min interval, the number of steps is equivalent to the hours ahead predicted. As expected, the *NMAE* and the *NRMSE* values increase as the number of steps increases, since the quality of the predictions shrinks with the prediction horizon.

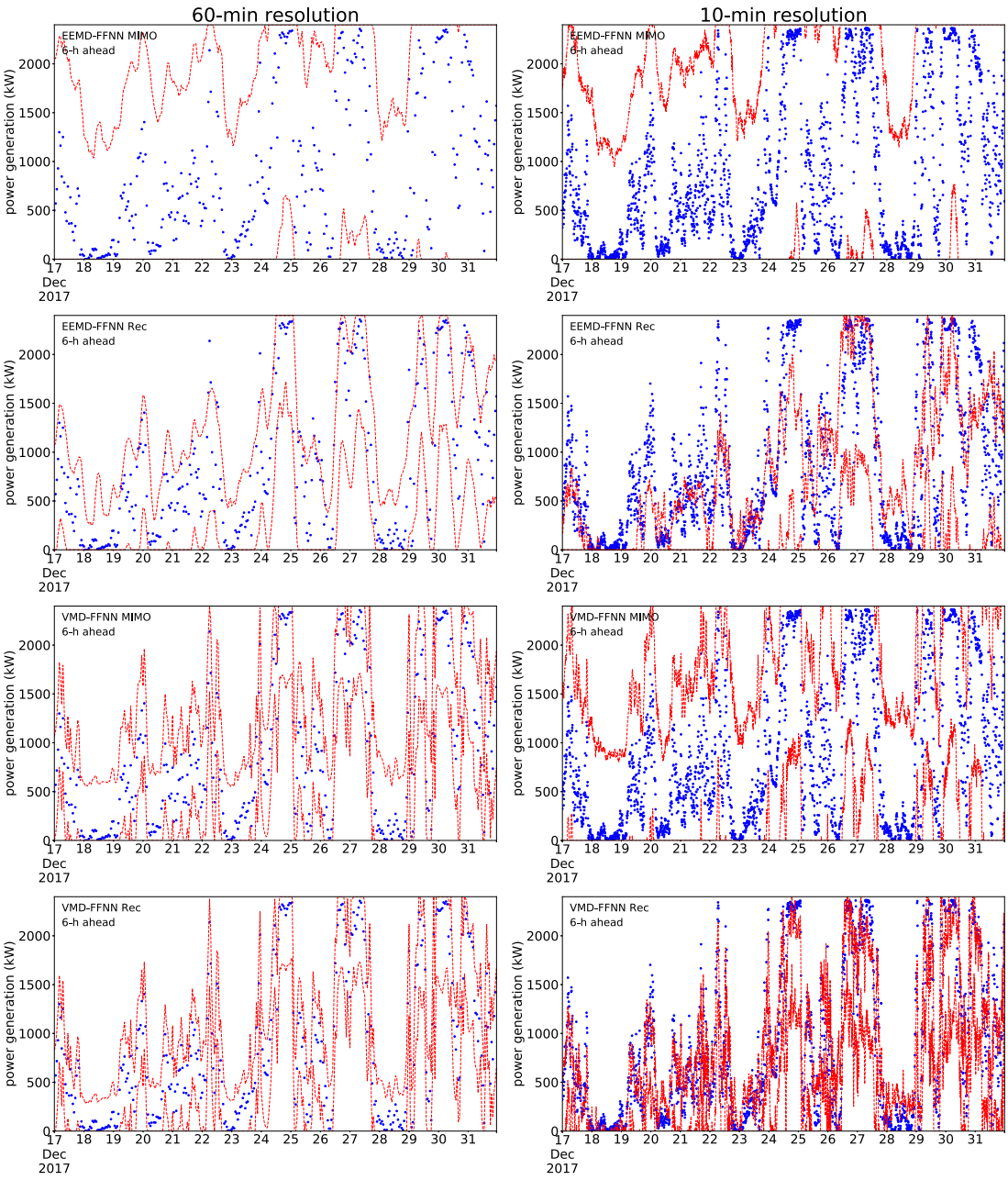

Fig. 9. 6-h ahead prediction intervals (95% confidence level) using the low-resolution (left) and high-resolution dataset (right).

The same expected behaviour is observed for the *IA*, indicating a lower performance for larger prediction horizons. The *NBIAS* is not affected by the prediction horizon, producing approximately regular values for the VMD-FFNN MIMO model and high variance for the other three models. The *SDE* tends to increase with larger prediction horizons. *MAPE* scores are higher for a larger number of steps, although this relationship is not linear.

Considering the techniques and available dataset, the VMD-FFNN MIMO model performs overall better than the rest of the models, and is used hereafter to discuss its robustness by analysing the daily values of the metrics obtained with the forecasts provided by this model (Fig. 8). Ideally, the forecast accuracy of the model should be as independent as possible from the data used to benchmark its validity, and therefore the numerical values obtained by the metrics should be similar for every subinterval. This performance is achieved for *NMAE* and *NRMSE* for three of the scenarios: 6-h ahead forecasts (both low- and high-resolution sets) and 24-h ahead forecasts with the low-resolution set. The lack of robustness for the other case (24-h ahead forecasts in the high-resolution set) results from a combination of several aspects: the model itself, the dataset, the prediction horizon length and the time-resolution of the data. Therefore, the model should be further calibrated to verify if the model is able to produce accurate forecasts for this case. The daily-averaged *NBIAS* shows low variability in the whole testing set for 6-h ahead forecasts, whereas they do not follow any pattern for 24-h ahead forecasts. The *NSDE* has a similar behaviour as the *NMAE* and *NRMSE*, producing robust outcomes for every case but 24-h ahead forecasts in the high-resolution dataset. The *IA* produces similar scores for 6-h ahead predictions in the low-resolution dataset. The results are quite steady for 6-h ahead (high-resolution set) and 24-h ahead (low-resolution set) forecasts, except for day 2, where there is a sudden drop in the performance of the metric. *MAPE* shows a great variability in its values, indicating a large sensitivity to changes of wind power output.

The performance evaluation metrics provide not only information about the general performance of the models in terms of accuracy,

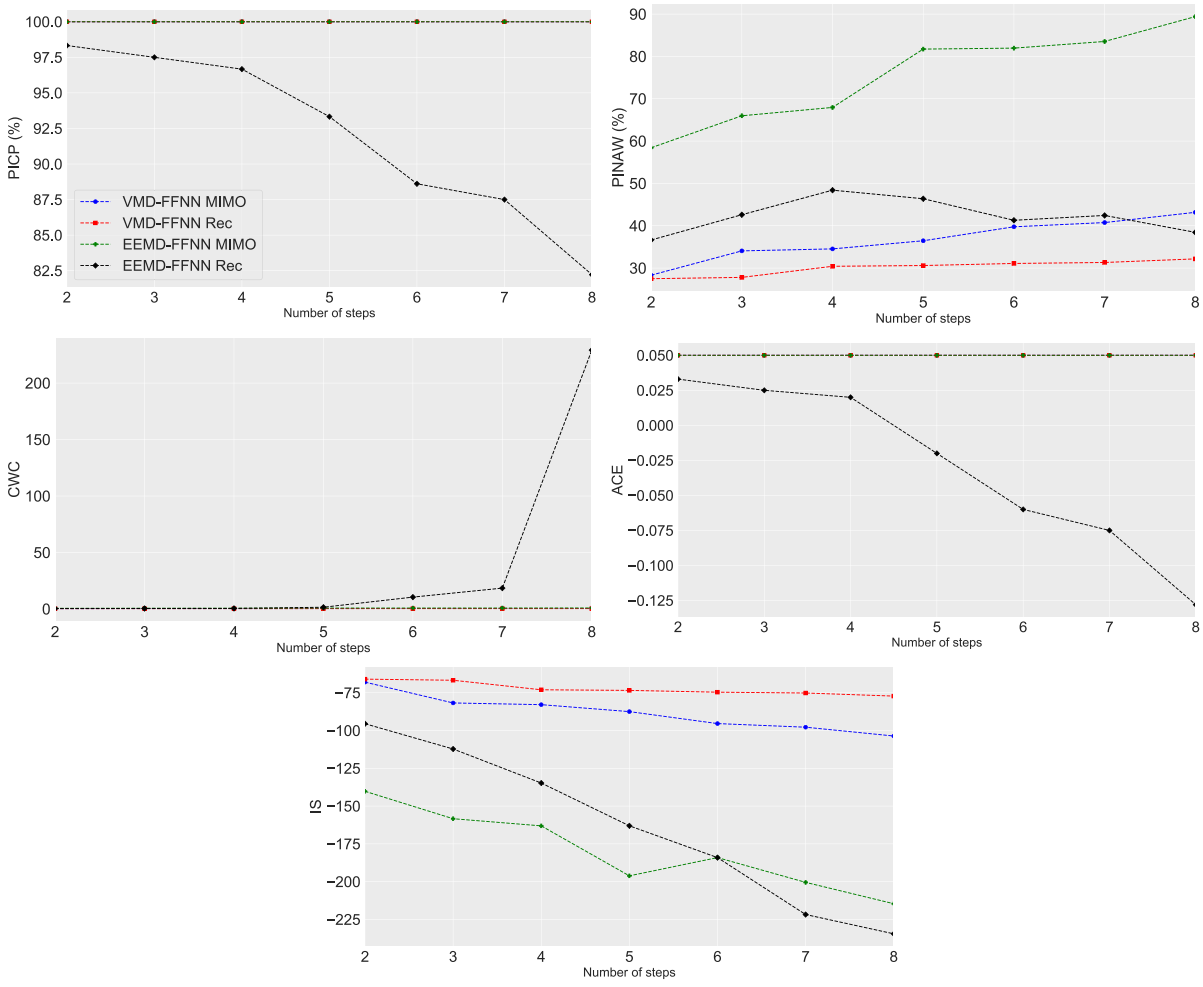

Fig. 10. Performance evaluation for interval metrics with respect to the number of steps ahead (low resolution dataset).

but also can act as a tool to analyse the effects of aspects such as the robustness of the models under different conditions. The VMD-FFNN MIMO model show a robust behaviour when the forecast accuracy is high since it shows low variations when calculating their values for different periods in the testing set. As expected, the performance evaluation metrics show a decreasing forecast accuracy while the prediction horizon increases. However, the *NBIAS* does not seem to be affected by the prediction horizon and keeps values around zero for the VMD-FFNN MIMO model. Additionally, the *MAPE* values are not in line with the scores obtained by the *NMAE*, the *NRMSE*, and the *IA* in terms of accuracy and consequently its use is not recommended.

5.3.2. Prediction intervals

PIs are estimated 6-h ahead for three confidence levels: 99%, 95%, and 80%. The results are shown in Table 6. As done previously for deterministic estimates, two FFNN models are used to benchmark the skill of the decomposition-based hybrid models.

The *PICP* and *ACE* measure exclusively the coverage of the PI. Considering that, these metrics indicate that the VMD-FFNN MIMO model presents the best results in terms of coverage, meaning that a larger number of observations fall within the interval. The *PINAW* and the *IS* quantify the width of the interval. The first metric only considers the width of the interval in every time step, while the *IS* penalizes incorrect intervals as shown in Eq. (16). In this study, narrower intervals are usually built when the recursive strategy is applied, at the expense of reducing the coverage of the interval. Therefore, better scores for *PINAW* are obtained for narrower intervals (60.05%, 31.09% and 14.26% for the VMD-FFNN Recursive model given 99%, 95%, and 80% confidence levels respectively). Furthermore, since the coverage of the intervals given by this model is high for all confidence levels, it also has the best scores for the *IS*. The *CWC* takes into account both *PICP* and *PINAW* to consider both features simultaneously. As stated in Eq. (14), those intervals where the *PICP* is lower than μ will be penalized and will produce equal measurements as the *PINAW*. Looking at the metrics' scores, the decomposition-based hybrid models present more skilled intervals that the FFNN models. Additionally, even if the EEMD-FFNN MIMO model has a high *PICP*, the intervals are not informative as they are also very wide, and therefore not useful for applications in the industry.

For the high-resolution dataset, the metrics also provide information regarding the multi-step ahead forecast strategy used. For instance, even if the *PINAW* is only 26.06% for the VMD-FFNN Recursive model compared to a 68.49% when the MIMO strategy is used (95% confidence level), the interval covers only a 57.78% of the data points of the testing set. The *IS* provides additional details in terms of the coverage-width relation: the VMD-FFNN MIMO model scores better in every scenario, meaning that this relation is balanced, as fewer intervals are penalized by not covering the actual values. In terms of skill, the behaviour is identical as the observed for the low-resolution dataset.

Fig. 9 displays the PIs obtained for the four models using the low-resolution and high-resolution set respectively. As the number of steps to predict is lower for the low-resolution dataset, intervals are more accurate and the boundaries are closer to the observations. In the low-resolution dataset, all models can produce reliable intervals, although the intervals are too wide for some of the models.

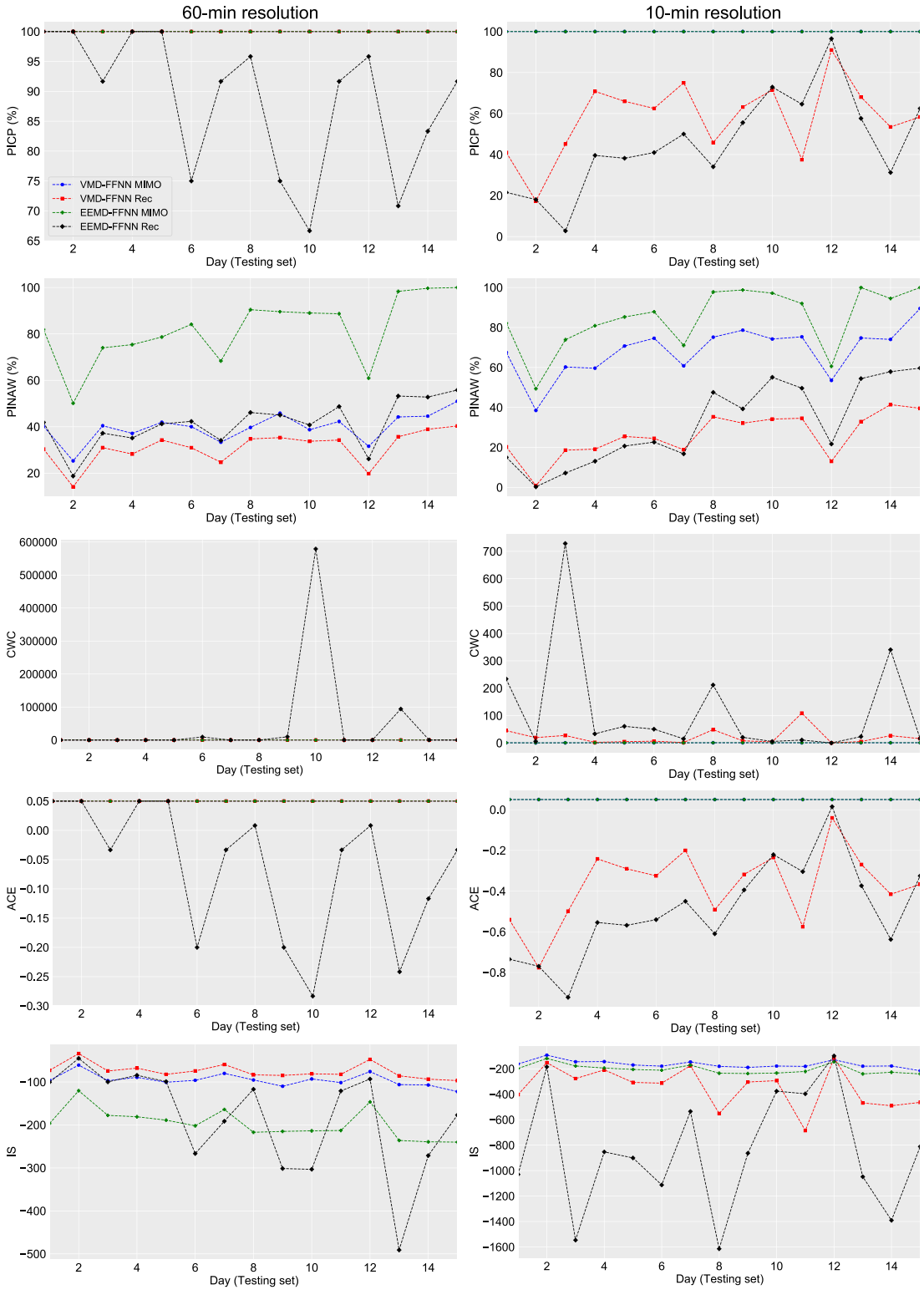

Fig. 11. Daily performance of metrics for PIs (95% confidence level).

The evolution of the PI performance metrics with respect to the prediction horizon length is shown in Fig. 10. The coverage of the PI decreases with the number of steps only for EEMD-FFNN Recursive model since the rest maintain all the values within the interval. This behaviour is replicated for the *ACE*, where the value decreases with a larger number of steps. The *PINAW* is barely influenced by the increase of the prediction horizon and it does not reveal high changes except for one of the models. Lastly, the *IS* decreases with the forecast horizon, indicating a lower forecast accuracy for longer horizons.

Fig. 11 shows the daily performance of the interval metrics for both sets. The *PICP* has reached its maximum value for some of the models, meaning that all the observations fall inside the interval. Therefore, the daily-averaged values of the *PICP* indicate that the models are robust in these cases. Otherwise, the metric shows the sensitivity of the models

to changes in the data as the *PICP* acts as a control depending on the percentage of values within the PI. The second metric (*PINAW*) shows a more robust performance for the models in the testing set, although it generates narrower intervals in days 2, 7 and 12 for every model. This behaviour takes place with low wind power generation, so these values seem justified as the interval will not grow any longer in its lower boundary. The *CWC* shows spikes whenever the *PICP* is lower than the parameter μ (set to the confidence level). It provides a control system to know the coverage of the interval and to what degree is good, as a larger *CWC* implies a greater penalization by the metric. However, as the *CWC* is very sensitive to this penalization, evaluating the robustness of the models with this metric is not advised. The *ACE* provides similar results than the *PICP*, only it takes into account the significance level. For this reason, except for exceptional cases, it is enough to choose only one of them to analyse the robustness of the interval. The daily-averaged *IS* produces robust values with low variability for the VMD-FFNN models in the low-resolution dataset. Additionally, the *IS* provides supplementary information about the forecast accuracy of the interval, since the prediction error by itself accounts only for the size of the PI. The more negative is the IS for a time step, the further away the actual value is from the interval.

The *PICP* and *PINAW* provide direct knowledge in terms of coverage and width. The *CWC* is highly sensitive to the coverage of the interval and the confidence level, and therefore not suitable to evaluate the robustness of the models. The *ACE* provides very similar information as the *PICP*, so its assessment is not necessary if the *PICP* is already estimated. The *IS* allows to determine reliably the robustness of a model, and provides information about the width interval while penalizing incorrect PIs in terms of coverage.

6. Conclusions

This paper presents an overview of the most common metrics applied for evaluating WPF models in the recent literature for deterministic and probabilistic estimates. Furthermore, this paper illustrates the capability of these metrics to properly evaluate the performance of WPF models over different datasets, time-resolution and other model specific attributes. This aspect is often disregarded to determine the validity of a forecasting model over an out-of-sample set, as the values of these metrics could be influenced by the intrinsic characteristics of the dataset and they could fluctuate considerably for different periods of the same testing set. A numerical study is presented using wind data from Ireland with two different time-resolutions (10-min and 60-min) and decomposition-based hybrid models to be assessed by the performance evaluation metrics.

Metrics are based on fundamental theory of statistics and can be considered robust as such. However, they capture different aspects of model performance. Most of the performance evaluation metrics identified for the assessment of deterministic estimates analyse all sources of error together (*MAE*, *RMSE*, *MAPE*, *IA*), while others evaluate a specific source of error such as the *BIAS*, that accounts for the systematic component of error, or the *SDE*, where only the random error is analysed. Probabilistic estimates account for both accuracy and precision, therefore they are preferred for model comparability. Their metrics evaluate the coverage provided by the interval, such as the *PICP*, or the width, such as the *PINAW*, whereas the *IS* provides additional information in terms of the overall quality of the interval. These three metrics give enough information to address the forecast accuracy of the interval. On the other hand, the *ACE* does not provide any additional information if the *PICP* is already estimated, consequently is not deemed necessary to evaluate the coverage of the interval. The *CWC* is highly sensitive to both the tuning parameters and the nature of the training set. Therefore, this metric is recommended as a parameter to train the data in methodologies such as the LUBE method, but not to evaluate the accuracy of a forecasting model.

The different performance evaluation methods are also considered to evaluate the robustness of the models over the testing set. Several aspects are consider such as the prediction horizon and the time-resolution of the data, as intra-hour wind data shows higher volatility. For higher resolution sets the models can be further calibrated to provide more accurate forecasts.

This study proves to be useful in the model development stage and aims to improve the benchmarking of WPF models by considering aspects such as the data, the prediction horizon, and the time-resolution, as well as the robustness of the models that can be evaluated through the use of the appropriate performance evaluation metrics. In addition, 6-h and 24-h ahead forecasts have been stressed since these horizons are relevant for wind power trading in electricity markets.

CRediT authorship contribution statement

J.M. González-Sopeña: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **V. Pakrashi:** Conceptualization, Validation, Investigation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition. **B. Ghosh:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation , Writing - original draft, Writing - review & editing, Supervision, Project administration, 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.

Acknowledgements

The authors acknowledge the funding of SEAI WindPearl, Ireland Project 18/RDD/263. Vikram Pakrashi would like to acknowledge the support of SFI MaREI centre and UCD Energy Institute, Ireland.

References

[1] Pinson P, Chevallier C, Kariniotakis GN. Trading wind generation from short-term probabilistic forecasts of wind power. IEEE Trans Power Syst 2007;22(3):1148–56.
[2] Bessa RJ, Matos MA, Costa IC, Bremermann L, Franchin IG, Pestana R, et al. Reserve setting and steady-state security assessment using wind power uncertainty forecast: A case study. IEEE Trans Sustain Energy 2012;3(4):827–36.
[3] Costa A, Crespo A, Navarro J, Lizcano G, Madsen H, Feitosa E. A review on the young history of the wind power short-term prediction. Renew Sustain Energy Rev 2008;12:1725–44.
[4] Foley AM, Leahy PG, Marvuglia A, McKeogh EJ. Current methods and advances in forecasting of wind power generation. Renew Energy 2012;37:1–8.
[5] Colak I, Sagiroglu S, Yesilbudak M. Data mining and wind power prediction: A literature review. Renew Energy 2012;46:241–7.
[6] Zhang Y, Wang J, Wang X. Review on probabilistic forecasting of wind power generation. Renew Sustain Energy Rev 2014;32:255–70.
[7] Bremnes JB. Probabilistic wind power forecasts using local quantile regression. Wind Energy: Int J Prog Appl Wind Power Convers Technol 2004;7(1):47–54.
[8] Pinson P, Kariniotakis G. On-line assessment of prediction risk for wind power production forecasts. Wind Energy Int J for Prog Appl Wind Power Convers Technol 2004;7(2):119–32.
[9] Pinson P, Madsen H, Nielsen HA, Papaefthymiou G, Klöckl B. From probabilistic forecasts to statistical scenarios of short-term wind power production. Wind Energy Int J for Prog Appl Wind Power Convers Technol 2009;12(1):51–62.
[10] Papaefthymiou G, Pinson P. Modeling of spatial dependence in wind power forecast uncertainty. In: Proc. 10th int. conf. probab. methods appl. to power syst.. IEEE; 2008, p. 1–9.
[11] Tascikaraoglu A, Uzunoglu M. A review of combined approaches for prediction of short-term wind speed and power. Renew Sustain Energy Rev 2014;34:243–54.
[12] Gallego-Castillo C, Cuerva-Tejero A, Lopez-Garcia O. A review on the recent history of wind power ramp forecasting. Renew Sustain Energy Rev 2015;52:1148–57.
[13] Yan J, Liu Y, Han S, Wang Y, Feng S. Reviews on uncertainty analysis of wind power forecasting. Renew Sustain Energy Rev 2015;52:1322–30.

[14] Chen C. Robustness properties of some forecasting methods for seasonal time series: a Monte Carlo study. Int J Forecast 1997;13(2):269–80.
[15] Sorensen P, Cutululis NA, Vigueras-Rodríguez A, Jensen LE, Hjerrild J, Donovan MH, et al. Power fluctuations from large wind farms. IEEE Trans Power Syst 2007;22(3):958–65.
[16] Hong T, Pinson P, Fan S. Global energy forecasting competition 2012. Int J Forecast 2014;30:357–63.
[17] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. Int J Forecast 2016;32:896–913.
[18] Giebel G, Kariniotakis G. Wind power forecasting—A review of the state of the art. In: Renewable energy forecasting. Elsevier; 2017, p. 59–109.
[19] Sweeney C, Bessa RJ, Browell J, Pinson P. The future of forecasting for renewable energy. Wiley Interdiscip Rev Energy Environ 2020;9(2):e365.
[20] Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS One 2018;13(3):e0194889.
[21] Dowell J, Pinson P. Very-short-term probabilistic wind power forecasts by sparse vector autoregression. IEEE Trans Smart Grid 2015;7(2):763–70.
[22] Messner JW, Pinson P. Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. Int J Forecast 2019;35(4):1485–98.
[23] Granger CW, Joyeux R. An introduction to long-memory time series models and fractional differencing. J Time Ser Anal 1980;1(1):15–29.
[24] Yuan X, Tan Q, Lei X, Yuan Y, Wu X. Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine. Energy 2017;129:122–37.
[25] Ozkan MB, Karagoz P. A novel wind power forecast model: Statistical hybrid wind power forecast technique (SHWIP). IEEE Trans Ind Inform 2015;11(2):375–87.
[26] Xie W, Zhang P, Chen R, Zhou Z. A nonparametric Bayesian framework for short-term wind power probabilistic forecast. IEEE Trans Power Syst 2018;34(1):371–9.
[27] Jiang Y, Xingying C, Kun YU, Yingchen L. Short-term wind power forecasting using hybrid method based on enhanced boosting algorithm. J Mod Power Syst Clean Energy 2017;5(1):126–33.
[28] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.
[29] Chang GW, Lu HJ, Chang YR, Lee YD. An improved neural network-based approach for short-term wind speed and power forecast. Renew Energy 2017;105:301–11.
[30] Chitsaz H, Amjady N, Zareipour H. Wind power forecast using wavelet neural network trained by improved clonal selection algorithm. Energy Convers Manage 2015;89:588–98.
[31] Osório G, Matias JCO, Catalão JPS. Short-term wind power forecasting using adaptive neuro-fuzzy inference system combined with evolutionary particle swarm optimization, wavelet transform and mutual information. Renew Energy 2015;75:301–7.
[32] Zameer A, Arshad J, Khan A, Raja MAZ. Intelligent and robust prediction of short term wind power using genetic programming based ensemble of neural networks. Energy Convers Manage 2017;134:361–72.
[33] Azimi R, Ghofrani M, Ghayekhloo M. A hybrid wind power forecasting model based on data mining and wavelets analysis. Energy Convers Manage 2016;127:208–25.
[34] Li S, Wang P, Goel L. Wind power forecasting using neural network ensembles with feature selection. IEEE Trans Sustain Energy 2015;6(4):1447–56.
[35] Qian Z, Pei Y, Zareipour H, Chen N. A review and discussion of decomposition-based hybrid models for wind energy forecasting applications. Appl Energy 2019;235:939–53.
[36] Naik J, Dash S, Dash PK, Bisoi R. Short term wind power forecasting using hybrid variational mode decomposition and multi-kernel regularized pseudo inverse neural network. Renew Energy 2018;118:180–212.
[37] Khosravi A, Nahavandi S, Creighton D. Prediction intervals for short-term wind farm power generation forecasts. IEEE Trans Sustain Energy 2013;4(3):602–10.
[38] Quan H, Srinivasan D, Khosravi A. Short-term load and wind power forecasting using neural network-based prediction intervals. IEEE Trans Neural Netw Learn Syst 2013;25(2):303–15.
[39] Kavousi-Fard A, Khosravi A, Nahavandi S. A new fuzzy-based combined prediction interval for wind power forecasting. IEEE Trans Power Syst 2015;31(1):18–26.
[40] Zou W, Li C, Chen P. An inter type-2 FCR algorithm based TS fuzzy model for short-term wind power interval prediction. IEEE Trans Ind Inform 2019;15(9):4934–43.
[41] Haque AU, Nehrir MH, Mandal P. A hybrid intelligent model for deterministic and quantile regression approach for probabilistic wind power forecasting. IEEE Trans Power Syst 2014;29(4):1663–72.
[42] He Y, Li H. Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. Energy Convers Manage 2018;164:374–84.
[43] Zhao Y, Ye L, Li Z, Song X, Lang Y, Su J. A novel bidirectional mechanism based on time series model for wind power forecasting. Appl Energy 2016;177:793–803.
[44] Hao Y, Tian C. A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. Appl Energy 2019;238:368–83.
[45] Zhang G, Wu Y, Wong KP, Xu Z, Dong ZY, Iu HH-C. An advanced approach for construction of optimal wind power prediction intervals. IEEE Trans Power Syst 2014;30(5):2706–15.
[46] Wan C, Lin J, Wang J, Song Y, Dong ZY. Direct quantile regression for nonparametric probabilistic forecasting of wind power generation. IEEE Trans Power Syst 2016;32(4):2767–78.
[47] Mahmoud T, Dong ZY, Ma J. An advanced approach for optimal wind power generation prediction intervals by using self-adaptive evolutionary extreme learning machine. Renew Energy 2018;126:254–69.
[48] Afshari-Igder M, Niknam T, Khooban M-H. Probabilistic wind power forecasting using a novel hybrid intelligent method. Neural Comput Appl 2018;30(2):473–85.
[49] Buhan S, Cadirci I. Multistage wind-electric power forecast by using a combination of advanced statistical methods. IEEE Trans Ind Inform 2015;11(5):1231–42.
[50] Liu J, Wang X, Lu Y. A novel hybrid methodology for short-term wind power forecasting based on adaptive neuro-fuzzy inference system. Renew Energy 2017;103:620–9.
[51] Lee D, Baldick R. Short-term wind power ensemble prediction based on Gaussian processes and neural networks. IEEE Trans Smart Grid 2013;5(1):501–10.
[52] Heinermann J, Kramer O. Machine learning ensembles for wind power prediction. Renew Energy 2016;89:671–9.
[53] Lahouar A, Slama JBH. Hour-ahead wind power forecast based on random forests. Renew Energy 2017;109:529–41.
[54] Wang H, Lei Z, Zhang X, Zhou B, Peng J. A review of deep learning for renewable energy forecasting. Energy Convers Manage 2019;198:111799.
[55] Wang H-z, Li G-q, Wang G-b, Peng J-c, Jiang H, Liu Y-t. Deep learning based ensemble approach for probabilistic wind power forecasting. Appl Energy 2017;188:56–70.
[56] Shi Z, Liang H, Dinavahi V. Direct interval forecast of uncertain wind power based on recurrent neural networks. IEEE Trans Sustain Energy 2017;9(3):1177–87.
[57] Qureshi AS, Khan A, Zameer A, Usman A. Wind power prediction using deep neural network based meta regression and transfer learning. Appl Soft Comput 2017;58:742–55.
[58] Zhang Y, Wang J. K-nearest neighbors and a kernel density estimator for gefcom2014 probabilistic wind power forecasting. Int J Forecast 2016;32(3):1074–80.
[59] Yan J, Li K, Bai E, Yang Z, Foley A. Time series wind power forecasting based on variant Gaussian process and TLBO. Neurocomputing 2016;189:135–44.
[60] Yang L, He M, Zhang J, Vittal V. Support-vector-machine-enhanced markov model for short-term wind power forecast. IEEE Trans Sustain Energy 2015;6(3):791–9.
[61] Wang Y, Hu Q, Meng D, Zhu P. Deterministic and probabilistic wind power forecasting using a variational Bayesian-based adaptive robust multi-kernel regression model. Appl Energy 2017;208:1097–112.
[62] Han L, Romero CE, Yao Z. Wind power forecasting based on principle component phase space reconstruction. Renew Energy 2015;81:737–44.
[63] Zjavka L, Mišák S. Direct wind power forecasting using a polynomial decomposition of the general differential equation. IEEE Trans Sustain Energy 2018;9(4):1529–39.
[64] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci Model Dev 2014;7(3):1247–50.
[65] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast 2006;22(4):679–88.
[66] Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 2016;32(3):669–79.
[67] Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS. Standardizing the performance evaluation of short-term wind power prediction models. Wind Eng 2005;29(6):475–89.
[68] Willmott CJ. On the validation of models. Phys Geogr 1981;2(2):184–94.
[69] Willmott CJ, Robeson SM, Matsuura K. A refined index of model performance. Int J Climatol 2012;32(13):2088–94.
[70] Gallego-Castillo C, Bessa R, Cavalcante L, Lopez-Garcia O. On-line quantile regression in the RKHS (Reproducing Kernel Hilbert Space) for operational probabilistic forecasting of wind power. Energy 2016;113:355–65.
[71] Lin Y, Yang M, Wan C, Wang J, Song Y. A multi-model combination approach for probabilistic wind power forecasting. IEEE Trans Sustain Energy 2018;10(1):226–37.
[72] Khorramdel B, Chung CY, Safari N, Price GCD. A fuzzy adaptive probabilistic wind power prediction framework using diffusion kernel density estimators. IEEE Trans Power Syst 2018;33(6):7109–21.
[73] Alessandrini S, Delle Monache L, Sperati S, Nissen JN. A novel application of an analog ensemble for short-term wind power forecasting. Renew Energy 2015;76:768–81.

[74] Pinson P, Kariniotakis G. Conditional prediction intervals of wind power generation. IEEE Trans Power Syst 2010;25(4):1845–56.
[75] Winkler RL. A decision-theoretic approach to interval estimation. J Am Stat Assoc 1972;67(337):187–91.
[76] Pinson P, Nielsen HA, Møller JK, Madsen H, Kariniotakis GN. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy Int J for Prog Appl Wind Power Convers Technol 2007;10(6):497–516.
[77] Pinson P. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. J R Stat Soc C 2012;61(4):555–76.
[78] Yan J, Liu Y, Han S, Qiu M. Wind power grouping forecasts and its uncertainty analysis using optimized relevance vector machine. Renew Sustain Energy Rev 2013;27:613–21.
[79] Zhang J, Meng H, Gu B, Li P. Research on short-term wind power combined forecasting and its Gaussian cloud uncertainty to support the integration of renewables and EVs. Renew Energy 2020;153:884–99.
[80] Soman SS, Zareipour H, Malik O, Mandal P. A review of wind power and wind speed forecasting methods with different time horizons. In: North American Power Symposium 2010. IEEE; 2010, p. 1–8.
[81] Taieb SB, Bontempi G, Atiya AF, Sorjamaa A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst Appl 2012;39(8):7067–83.
[82] Wang J, Song Y, Liu F, Hou R. Analysis and application of forecasting models in wind power integration: A review of multi-step-ahead wind speed forecasting models. Renew Sustain Energy Rev 2016;60:960–81.
[83] Taieb SB, Sorjamaa A, Bontempi G. Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing 2010;73(10–12):1950–7.
[84] Sorjamaa A, Lendasse A. Time series prediction using DirRec strategy. In: Eur. symp. artif. neural networks, Vol. 6. 2006, p. 143–8.
[85] Taieb SB, Bontempi G, Sorjamaa A, Lendasse A. Long-term prediction of time series by combining direct and MIMO strategies. In: 2009 int. jt. conf. neural networks. IEEE; 2009, p. 3054–61.
[86] Wu Z, Huang NE. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 2009;1(01):1–41.
[87] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond 1998;454(1971):903–95.
[88] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process 2013;62(3):531–44.
[89] Hestenes MR. Multiplier and gradient methods. J Optim Theory Appl 1969;4(5):303–20.
[90] Koenker R, Bassett Jr G. Regression quantiles. Econometrica 1978;46(1):33–50.
[91] Cannon AJ. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Comput Geosci 2011;37(9):1277–84.
[92] Nielsen HA, Madsen H, Nielsen TS. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. Wind Energy: Int J Prog Appl Wind Power Convers Technol 2006;9(1–2):95–108.