Technology and Economics of Smart Grids and Sustainable Energy (2020) 5:19
https://doi.org/10.1007/s40866-020-00090-8

ORIGINAL PAPER

# Robust Assessment of Short-Term Wind Power Forecasting Models on Multiple Time Horizons

Fabrizio De Caro[1] · Jacopo De Stefani[2] · Gianluca Bontempi[2] · Alfredo Vaccaro[1] · Domenico Villacci[1]

Received: 7 February 2020 / Accepted: 24 August 2020


## Abstract

The massive penetration of renewable power generation in modern power grids is an effective way to reduce the impact of energy production on global warming. Unfortunately, the wind power generation may affect the regular operation of electrical systems, due to the stochastic and intermittent nature of the wind. For this reason, reducing the uncertainty about the wind evolution, e.g. by using short-term wind power forecasting methodologies, is a priority for system operators and wind producers to implement low-carbon power grids. Unfortunately, though the complexity of this task implies the comparison of several alternative forecasting methodologies and dimensionality reduction techniques, a general and robust procedure of model assessment still lacks in literature. In this paper the authors propose a robust methodology, based on extensive statistical analysis and resampling routines, to supply the most effective wind power forecasting method by testing a vast ensemble of methodologies over multiple time-scales and on a real case study. Experimental results on real data collected in an Italian wind farm show the potential of ensemble approaches integrating both statistical and machine learning methods.

**Keywords** Wind Power Forecasting · Wind Energy · Robust Forecasting · Ensemble Forecasting

## Nomenclature

### Symbols

| | |
|---|---|
| $\mathbf{D}[N, S]$ | Matrix of observations of size $[N, S]$ |
| $\mathbf{P}[N, \phi]$ | Predictor matrix of size $[N, \phi]$ |
| $\mathbf{R}[N, \rho]$ | Target matrix of size $[N, \rho]$ |
| $N$ | Number of observations |
| $\phi$ | Number of predictors |
| $\rho$ | Number of targets |
| $c$ | Number of lagged time windows |
| $\gamma$ | Number of smoothed variables |
| $L$ | Auto-regressive lag |
| $H$ | Forecasting horizon (number of steps ahead) |
| $\Phi$ | Number of features after pre-processing |
| $\mathbf{X}_0$ | Embedded Input Matrix |
| $\mathbf{Y}_0$ | Embedded Output Matrix |
| $\mathbf{X}^{(v)}$ | Matrix $\mathbf{X}$ for test case $v$ |
| $\mathbf{Y}^{(v)}$ | Matrix $\mathbf{Y}$ for test case $v$ |
| $\mathbf{X}^{(v)}_{trn}$ | Training matrix for test case $v$ |
| $\mathbf{X}^{(v)}_{val}$ | Validation matrix for test case $v$ |
| $r_{1,\ldots,N}$ | Generic smoothed time series |
| $RSS$ | Residual Sum of Squares |
| $TSS$ | Total Sum of Squares |
| $A$ | Number of samples in validation set |
| $y_a$, $\hat{y}_a$ $a^{th}$ | actual,predicted power in validation set |
| $\hat{y}$ | mean actual value in validation set |
| $\sigma$ | actual power Standard deviation in validation set |

### Abbreviations

| | |
|---|---|
| Ad. | Adaptive Ensemble Forecasting |

✉ Fabrizio De Caro
fdecaro@unisannio.it

Jacopo De Stefani
jacopo.de.stefani@ulb.ac.be

Gianluca Bontempi
gbonte@ulb.ac.be

Alfredo Vaccaro
vaccaro@unisannio.it

Domenico Villacci
villacci@unisannio.it

[1] University of Sannio, piazza Roma 21, Benevento, 82100, Italy

[2] Universitè Libre de Bruxelles, Campus de la Plaine ULB CP212, boulevard du Triomphe, 1050 Bruxelles, Belgium

Published online: 18 October 2020

| | |
|---|---|
| | (dynamic weights) |
| ANN | Artificial Neural Network |
| ARIMA | Auto Regressive Integrated Moving Average |
| ARMA | Auto Regressive Moving Average |
| CFD | Computational Fluid Dynamics |
| DEM | Digital Elevation Model |
| ELM | Extreme Learning Machine |
| GRU | Gated Recurrent Unit (ANN) |
| GBM | Gradient Boosting Machine |
| HW-ES | Holt-Winter Exponential Smoothing |
| HYB | Hybrid method |
| LSTM | Long Short Term Memory unit |
| M3 | M3 competition |
| ML | Machine Learning |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| MSE | Mean Square Error |
| mRMR | minimum Redundancy Maximum Relevancy |
| nMSE | Normalized MSE |
| NWP | Numerical Weather Prediction |
| PH | Physical method |
| PCA | Principal Component Analysis |
| RF | Random Forest |
| RMSE | Root Mean Square Error |
| $R^2$ | R-squared error |
| Av. | Simple Ensemble Forecasting |
| SCADA | Supervisory Control and Data Acquisition |
| SF | Statistical Forecasting |
| SVM | Support Vector Machine |
| WPF | Wind Power Forecasting |
| WTG | Wind Turbine Generator |

## Introduction

The wide scale employment of renewable generation plants in modern power grids is one of the most effective ways to produce electric energy without increasing the emission of greenhouse gases. [17]. However, the large penetration of renewable power generation may cause several issues in the operation of the electric grids, due to the stochastic and intermittent nature of the renewable sources [12, 33]. The management of uncertainty is then a major challenge for system operators [8, 16], and wind producers [39].

In particular, the industrial applications for wind power forecasting from 1 to 6 hours are strongly related to the wide penetration of renewable variable energies in modern power grids. This is pushing a transition phase in the electricity markets to become more flexible causing the introduction of rolling markets, as done in UK and Australia, and shorter intra-day-auction, as done in Europe, in order to correct the scheduling of day auction markets [45]

Particularly, the intra-day market platforms are structured as both discrete auction market and continuous intra-day markets, which operates with time ranges from 5 to 90 minutes. For this reason, the adoption of accurate wind power forecasting after the day-ahead gate closure allows the wind producers to tune up the bidding offers in intra-day markets trying to correct the previous forecasting error in advance, especially in the continuous market platform of tomorrow.

Hence, in according to the opportunity to adjust the bidding offers in continuous market platforms all the chain of wind power forecasting models over the time deserve of the same importance, needing high accuracy in according to the requested forecasting horizon.

Furthermore, transmission system operators uses short term wind power forecasting to manage imbalances and ancillary services procurement. Thus, in light of this developing a pipeline for WPFs model represent a strategic concept to chose the model that best suites the corresponding needs.

In literature, many statistical, physical and hybrid approaches have been proposed for Wind Power Forecasting (WPF) at different timescales and with different objectives. In particular, we distinguish between: a) ultra-short-term forecasting (1 minute - 1 hour) to address real-time activities in electricity market and wind turbine control, b) short term (1h - 6h) forecasting for balancing activities, c) medium-term (6h - 1 week) forecasting for energy market bidding and d) long-term forecasting (> 1w) for planning and maintenance of wind farms [49].

As far as statistical approaches are concerned, we distinguish between statistical forecasting (SF) and machine learning (ML). The former approach makes a number of explicit assumptions (e.g. parametric distribution, non-multicollinearity, homoscedasticity) about the data distribution while the second relies on non parametric and data-driven strategies for fitting and model selection.

Well-known examples of statistical forecasting are the Auto-Regressive Moving Average (ARMA) proposed in [42] for power generation prediction of wind farms and the Box-Jenkin method [25] based on fractional auto-regressive integrated moving average (f-ARIMA) models.

ML models make no explicit assumptions about data distribution and can be characterized in terms of the adopted hierarchical and multi-step-ahead strategy. In terms of hierarchical forecasting we distinguish between the spot approach returning the prediction for each wind generator (with detrimental impact on the computational cost) and the aggregate approach which targets the overall wind farm power generation (suitable for large area prediction and when producers share confidential data to system operators).

In terms of multi-step-ahead strategy [5] we distinguish here between the iterated strategy approach which uses only one trained model to predict the full forecasting span horizon and the direct approach which learns a different models for any step of the forecasting span.

While the former approach is easy to implement but is highly sensitive to the estimation error [10], the second does not take into consideration any relationship between predicted values at different horizons [4]. Furthermore, novel strategies aimed at considering the forecasting error as correction factor in multi-step ahead WPF have been in introduced by [29] and [21].

Several examples of ML solutions for wind forecasting have been proposed in literature: [48] proposed a WPF model based on Support Vector Machine (SVM), [43] used Deep Learning Models based on an Extreme Learning Machine (ELM) to supply Prediction Interval for a large wind farm and [2] introduced a Hybrid Neural Network architecture, [44] proposed a deep learning ensemble approach and walevet decomposition for probabilstic WPF.

Physical (PH) approaches supply the wind power forecasting by using mathematical models of the weather at large scale, called Numerical Weather Predictions (NWP), and adapting them to regional scale by using a Digital Elevation Model (DEM), which allows to consider terrain orography and roughness by solving Computational Fluid Dynamic problems (CFD). In a second phase, the wind speed is converted into power by using positions and power curves of the wind generators. Unfortunately, the high computational cost restrains their application to medium and long term forecasting horizons.

Hybrid (HYB) approaches aims to keep the best of both worlds by combining NWP with statistical post processing [18, 46].

Since the literature on WPF is wide, there is an increasing need for a quantitative and statistically founded comparison of existing approaches. Nevertheless little effort has been done so far in this direction. Among the few examples we cite [13], which tested several WPF models on different horizons in terms of percentage Mean Absolute Percentage Error (MAPE), [11] discussing a spatial comparison on daily wind power forecasting and [26] which took into consideration only the mean value for an ensemble of accuracy metrics.

Korprasertsak and Leephakpreeda [26] presents an assessment in terms of prediction interval of several Artificial Neural Networks models while [35] introduced the t-student test between the assessed models for short term WPF models. Unfortunately, the t-test is subject to $\alpha$ (probability error of I type) inflation in case of multiple comparisons [27], reducing its reliability.

For these reasons, we have proposed an exhaustive statistical analysis of many promising forecasting models, inspired to the assessment procedure effected by the M3 competition [32]. M3 has been the third edition of a famous forecasting competition where the competing models have been widely assessed using large number of data-sets, differing for nature and features, and a vast number of accuracy metrics.

This paper intends to bring the following main contributions to the literature:

1. The introduction of a reliable WPF benchmark, which processes the power generation, the wind speed and the direction profiles of a real 30MW wind farm (15 wind generators) situated on the ridge of Apennines in southern Italy, and characterized by a very complex orography
2. the formalization of a novel pipeline aimed at designing, and assessing a data driven WPF model;
3. a rigorous statistical assessment of different Statistical and Machine Learning-based forecasting models on several prediction horizons, ranging from 1 to 6 hours ahead;
4. the conceptualization of multiple dynamic techniques operating in ensemble prediction for solving the wind power forecasting problem;
5. the development of a novel tool for decision support strategy aimed at analyzing the performance of a large set of heterogeneous forecasting models on multiple case studies, and identifying, for each prediction horizon, the most reliable forecasting model;

The expectation of the authors are the results of this paper will support wind producers and system operators [1] in the choice of the most suitable model for their needs, in light of the opening of the capacity market to the not programmable renewable energies [39].

## A Modeling Pipeline for WPF

This section details the main steps of a pipeline for designing and assessing a WPF model from observed data: data filtering of raw data, feature engineering to enhance the information of the extracted signal, data embedding to enable the adoption of supervised learning models, feature selection to reduce dimensionality and resampling to carry out a robust performance analysis on a large ensemble of cases (Fig. 1).

### Filtering of Raw Data

The raw signals collected in a wind farm are typically the output of a Supervisory Control And Data Acquisition (SCADA) system that returns series with a time resolution of 10 minutes. This system equips all wind generators/anemometers of the wind farm and supplies

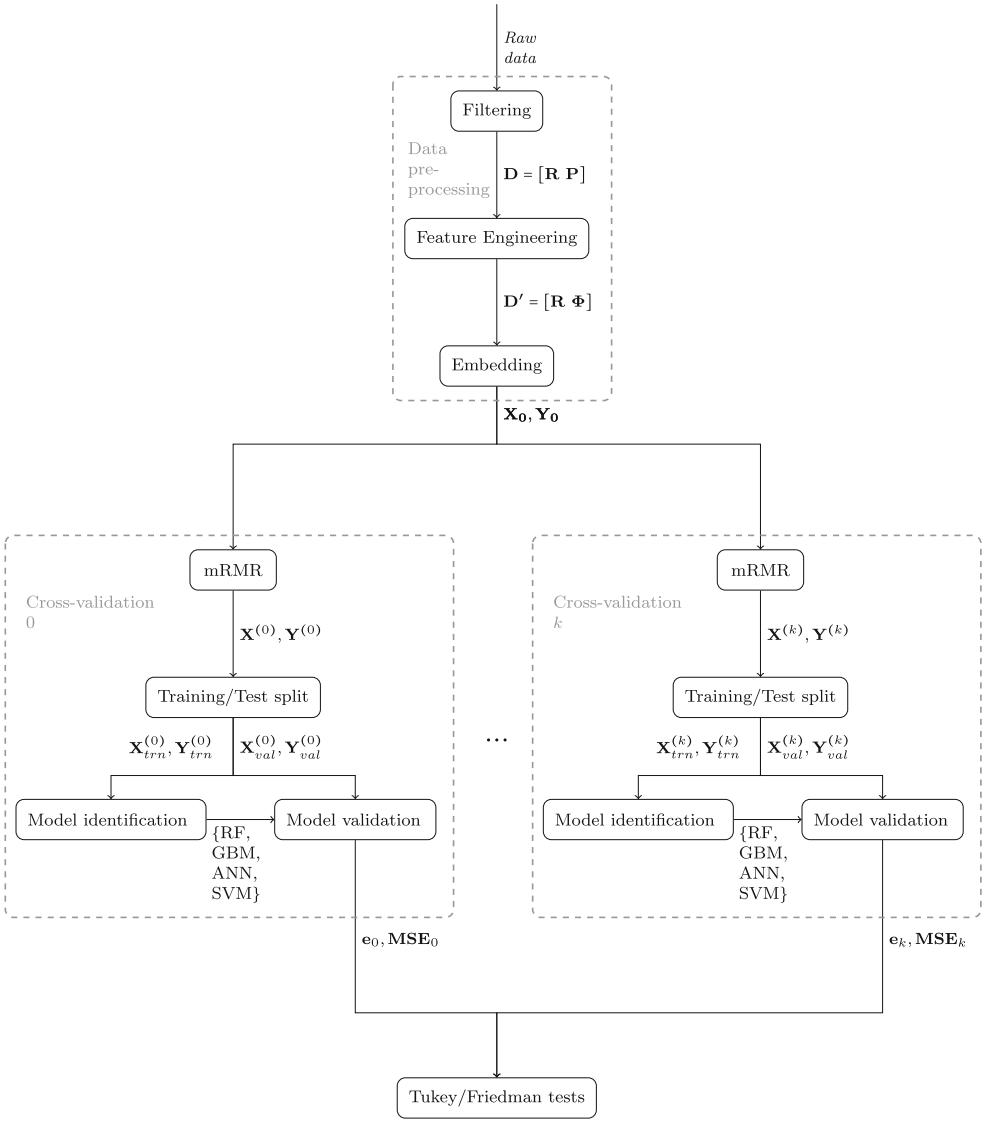


**Fig. 1** Flow chart of the proposed pipeline

both environmental and mechanical data. All data are radio broadcast to a central server where a database is continuously updated. The first data processing consists in extracting and grouping data according to the source (generator/anemometer). Though the total amount of data is larger, in this paper we will limit to power generation, temperature, measured wind speed/direction of the wind generators and wind speed/direction of the anemometers.

## Feature Engineering

In this step the raw signals are formatted into a matrix $\mathbf{D}[N, S]$, where $N$ and $S$ denote the number of samples and variables respectively. This matrix can be decomposed into the predictor submatrix $\mathbf{P}[N, \phi]$ and target submatrix $\mathbf{R}[N, \rho]$, where $\phi$ and $\rho$ stand for the number of predictors and targets, respectively. This decomposition is supposed to make easier the feature engineering step, which consists in augmenting the original dataset with a number of statistics related to the past behavior of the series.

$$\mathbf{D} = \big[\mathbf{R}\ \mathbf{P}\big] = \left[\begin{bmatrix} r_{11} & \dots & r_{1\rho} \\ \vdots & \dots & \vdots \\ r_{N1} & \dots & r_{N\rho} \end{bmatrix} \begin{bmatrix} p_{11} & \dots & p_{N\phi} \\ \vdots & \dots & \vdots \\ p_{N1} & \dots & p_{N\phi} \end{bmatrix}\right] \tag{1}$$

In particular, we consider here:

– *Moving Average*: arithmetic mean on the past $Q$ observations:

$$r_{s_{1,t}} = \frac{1}{Q} \sum_{i=0}^{i=Q-1} y_{(t-i)}, \quad \forall t \in [1, N] \tag{2}$$

– *Maximum Value*: local maximum value for the considered time windows:

$$r_{s_{2,t}} = \max\left(y_{(t)}, \ldots, y_{(Q-1)}\right), \quad \forall t \in [1, N] \tag{3}$$

– *Incremental Ratio of the Moving Average*: average incremental ratio over the time

$$r_{s_{3,t}} = \frac{1}{Q}\left[\left(\frac{1}{Q}\sum_{i=0}^{i=Q-1} y_{(t-i)}\right) - \left(\frac{1}{Q}\sum_{i=Q}^{i=2(Q-1)} y_{(t-i)}\right)\right], \quad \forall t \in [1, N] \tag{4}$$

– $p - quantile$: The p-quantile is the value, given a sortable observation set, for that the probability to have observations lying below this value is the $p\%$.

The feature creation leads to an increase of the size of $\mathbf{P}$ to $[N, \Phi]$, where $\Phi = (\phi + \rho c \gamma)$ and $c$ and $\gamma$ are the number of lagged time windows and smoothed variables, respectively.

## Embedding Procedure

This step rearranges the sub-matrices $\mathbf{P}$ and $\mathbf{R}$ in an embedded input/output form depending on the horizon $H$ and time lag $L$.

The outcome is made of the predictor matrix $\mathbf{X_0}$ in (5) and the target $\mathbf{Y_0}$ matrix (6) whose dimensions are $[N - L - H, \Phi \cdot L]$ and $[N - L - H, H]$, respectively. Note that the increment in size is due to the generation of as many variables as are the number of the time steps between the current time sample and the maximum lag for each variable.

$$\mathbf{X_0} = \begin{pmatrix} \overbrace{t-0 \;\; t-1 \;\; t-2 \;\; \cdots \;\; t-L+1}^{p_1} & \cdots & \overbrace{t-0 \;\; t-1 \;\; t-2 \;\; \cdots \;\; t-L+1}^{p_\Phi} \\ p_{11} \;\; - \;\; - \;\; \cdots \;\; - & \cdots & p_{1\Phi} \;\; - \;\; - \;\; \cdots \;\; - \\ p_{21} \;\; p_{11} \;\; - \;\; \cdots \;\; - & \cdots & p_{2\Phi} \;\; p_{1\Phi} \;\; - \;\; \cdots \;\; - \\ p_{31} \;\; p_{21} \;\; p_{11} \;\; \cdots \;\; - & \cdots & p_{3\Phi} \;\; p_{2\Phi} \;\; p_{1\Phi} \;\; \cdots \;\; - \\ p_{41} \;\; p_{31} \;\; p_{21} \;\; \cdots \;\; \vdots & \cdots & p_{4\Phi} \;\; p_{3\Phi} \;\; p_{2\Phi} \;\; \cdots \;\; - \\ \vdots \;\; \vdots \;\; \vdots \;\; \ddots \;\; \vdots & \cdots & \vdots \;\; \vdots \;\; \vdots \;\; \ddots \;\; \vdots \\ p_{N1} \;\; p_{(N-1)1} \;\; p_{(N-2)1} \;\; \cdots \;\; p_{(N+1-L)1} & \cdots & p_{N\Phi} \;\; p_{(N-1)\Phi} \;\; p_{(N-2)\Phi} \;\; \cdots \;\; p_{(N+1-L)\Phi} \end{pmatrix} \tag{5}$$

$$\mathbf{Y_0} = \begin{pmatrix} \overbrace{t+1 \;\; t+2 \;\; t+3 \;\; \cdots \;\; t+H}^{r_1} & \cdots & \overbrace{t+1 \;\; t+2 \;\; t+3 \;\; \cdots \;\; t+H}^{r_\rho} \\ r_{21} \;\; r_{31} \;\; r_{41} \;\; \cdots \;\; r_{(t+H)1} & \cdots & r_{2\rho} \;\; r_{3\rho} \;\; r_{4\rho} \;\; \cdots \;\; r_{(t+H)\rho} \\ r_{31} \;\; r_{41} \;\; r_{51} \;\; \cdots \;\; \vdots & \cdots & x_{3\rho} \;\; x_{4\rho} \;\; - \;\; \cdots \;\; \vdots \\ r_{41} \;\; r_{51} \;\; r_{61} \;\; \cdots \;\; \vdots & \cdots & x_{4\rho} \;\; x_{5\rho} \;\; x_{6\rho} \;\; \cdots \;\; \vdots \\ r_{51} \;\; r_{61} \;\; r_{71} \;\; \cdots \;\; \vdots & \cdots & r_{5\rho} \;\; r_{6\rho} \;\; r_{7S} \;\; \cdots \;\; \vdots \\ \vdots \;\; \vdots \;\; \vdots \;\; \ddots \;\; \vdots & \cdots & \vdots \;\; \vdots \;\; \vdots \;\; \ddots \;\; \vdots \\ r_{(N-H)1} \;\; - \;\; - \;\; \cdots \;\; - & \cdots & x_{(N-H)\rho} \;\; - \;\; - \;\; \cdots \;\; - \end{pmatrix} \tag{6}$$

Once the matrices $\mathbf{X_0}$ and $\mathbf{Y_0}$ are available, a number $V$ of training and test sets are defined in order to implement a rolling window assessment strategy [3].

In particular, for a generic case test $v$, the matrices $\mathbf{X}^{(v)}$ and $\mathbf{Y}^{(v)}$ are used to derive the training matrices $\mathbf{X}^{(v)}_{trn}$, $\mathbf{Y}^{(v)}_{trn}$ and the validation matrices $\mathbf{X}^{(v)}_{val}$, $\mathbf{Y}^{(v)}_{val}$, which will be used later to assess the prediction models.

## Feature Selection

Though feature engineering and data embedding generate useful information for forecasting, they cause a large increase in data dimension and a consequent number of drawbacks like: curse of dimensionality, high demand of computational resources and ill-conditioning in data analysis [28].

Hence, it is recommended to adoption of dimensionality reduction or features selection techniques [38]. Dimensionality reduction techniques combine original features to provide a smaller number of features with enhanced predictive power. A well-known example is the Principal Component Analysis (PCA) which creates by linear combination orthogonal variables expected to retain a large part of the data variance. Feature selection extracts a subset of variables expected to be as relevant as possible for the prediction target. Plenty of feature selection techniques have been proposed in literature. Here we will limit to consider an information-theoretic filter, called minimum Redundancy Maximum Relevancy (mRMR) [24]. The motivation of this choice is that this algorithm is a fast and effective way to select a number of features which are highly informative and low redundant. mRMR selects the features by maximizing the average mutual information and minimizing the redundancy according to the following incremental algorithm [20]:

$$\max_{x_j \in X - S_{m-1}} \left[ I(x_j; b) - \frac{1}{1-G} \sum\nolimits_{x_i \in S_{G-1}} I(x_i; x_j) \right] \tag{7}$$

where $S$ is the ensemble predictors, $x_j$ is the j-th analyzed predictor, $x_i$ is the i-th compared predictor to $x_j$, $G$ is the number of predictors, and $b$ a generic target. $I(x_j; b)$ and $I(x_i; x_j)$ are estimated by computing the mutual information between the corresponding couples of variables.

In particular, in this work mRMR has been chosen as technique to reduce the cardinality of the embedded predictor matrix after preliminary tests, which have highlighted its greater effectiveness respect with the PCA.

## Model Generation

The number of predictive models proposed in the data science literature is huge as shown in Table 1 we will limit to consider a small number of statistical, ML models and their combination (Ensemble).

**Table 1** Assessed models

| | |
|---|---|
| Ad. ANNs - Naïve | Ad. Naïve-RF-SVM-GBM |
| GRUdrop (ANN) | Ad. SVM-Naïve |
| GRU (ANN) | Ad. GBM-Naïve |
| Simple ANN | Ad. RF-Naïve |
| Exponential Smoothing | Av. SVM-GBM |
| SVM | Av. SVM-Naïve |
| Av. GBM-RF | GBM |
| Av. Naïve-RF | RF |
| Av. GBM-Naive | |

### Statistical Models

We consider two types of statistical models: (i) the Naive model returning a simple moving average mean on the past observed values and (ii) the Holt Winter Exponential Smoothing [22], which has proved its effectiveness in predicting complex time series [19] by decomposing the signal in trend, seasonal and random components in order to estimate independently the evolution of each of them.

### Machine Learning Models

The input/output representation of the data allows the adoption of generic supervised learning algorithms. Here we will focus on Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machines (SVM) and Artificial Neural Networks (ANNs).

RF relies on the idea that many low-correlated and unbiased weak predictors, which are characterized by not much great variance, can improve the prediction accuracy respect with the employment of a single predictor [30]. RF operates by producing many weak models (regression trees) from data subsets having different samples and features, which are extracted randomly from original training set in order to reduce the correlation between the weak predictors, where the final prediction will be the average between all weak predictions.

GBM, on the other hand, creates an ensemble by iterative addiction of weak predictors (regression trees), each one learning on the residuals from the previous ensemble [14].

Differently, SVM for regression works differently from the previous methods because it relies on minimizing a loss function by considering the only samples lying outside a tolerance bound, neglecting the others. Because the SVM optimization problem basically finds hyper-plane coefficients that satisfy a set of constraints, it performs a linear regression. In order to apply it to data with non-linear dependencies, a kernel function [6] is used to map the original space onto a higher dimensional space where the data is linearly separable.

When the relation on data are strong nonlinear ANN may be effective in predicting complex time series evolution. Particularly, many Recurrent Neural Networks have been compared here, where each of them has been equipped with different architecture. In particular, Long Short-Term Memory (LSTM) has been widely applied for wind and solar forecasting [31], whereas Gated Recurrent Units (GRU) units represents an evolution of LSTM to catch dependencies at different timescales [9].

Furthermore, the adoption of the dropout technique to GRU should avoid over-fitting issue, by increasing the forecasting accuracy as shown in [40].

### Ensemble Forecasting

Many previous models such as RF and GBM rely on the ensemble forecasting concept, where many low-correlated weak predictors are trained, providing the final prediction by aggregating all model outputs. Now, the same idea is applied at a higher level by combining the predictions of different models in a certain way, producing an ensemble averaging [34].

In particular, the ensemble forecast uses many forecasting models operating in parallel, with the aim at improving the forecasting accuracy respect to the performance supplied by each single method.

This because each forecasting model operates on different base assumptions, performing better/worse depending on many factors, such as the differences between characteristic of training and validation data sets, respect with the others causing a local change in performance over the time.

Therefore, the aggregation method plays a crucial role in combining the single predictors output where the most employed are the simple averaging (Av) and the adaptive weighted averaging (Ad). The latter considers dynamical weights that are updated regularly over the time, giving greater weight to models with lower RMSE as shown by (8).

$$x_j^{(t)} = \left(\mathrm{RMSE}_j^{(t)}\right)^{-1} \left(\sum_{k=1}^{K} \left(\mathrm{RMSE}_k^{(t)}\right)^{-1}\right)^{-1}, \quad x_j \in [0,1] \tag{8}$$

where $K$ is the number of considered machine learning models, whereas (9) shows how the joint forecasting is computed, where $f_k$ is k-th model of the forecasting ensemble.

$$\hat{y}_{(t+i)} = \sum_{k=1}^{K} f_{k(t+i)} x_k^{(t)}, \quad \forall i \in [1, H] \tag{9}$$

**Table 2** Metrics

| metric | equation | domain |
|---|---|---|
| MSE | $\frac{1}{A}\sum_{a=1}^{Z}\left(\hat{y}_a - y_a\right)^2$ | $[0, \infty]$ |
| MAE | $\frac{1}{A}\sum_{a=1}^{A}\left\|\hat{y}_a - y_a\right\|$ | $[0, \infty]$ |
| $R^2$ | $1 - \frac{\text{RSS}}{\text{TSS}}$<br>$\text{RSS} = \sum_{a=1}^{A}\left(\hat{y}_a - y_a\right)^2$<br>$\text{TSS} = \sum_{a=1}^{A}\left(y_a - \bar{y}\right)^2$ | $[0, 1]$ |
| nMSE | $\frac{\text{MSE}}{\sigma}$ | $[0, \infty]$ |

$A$ is the number of samples in validation dataset

$\sigma$ is the variance of validation dataset

## Model Validation

The choice of the most suitable metric ([37]) to assess wind power forecasting is still an open issue [45]. Here we adopt the metrics detailed in Table 2 i.e. Mean Square Error (MSE), Mean Absolute Error (MAE) and R squared ($R^2$) [7].

In order to robustly assess the alternative WPF models, a large number of cases ($V = 17$) has been considered. Each case is based on different pairs of training and validation sets [23], generated by application of the rolling window technique [41].

The aggregation of the resulting accuracy measures enables a thorough assessment of each WPF model by means of the Friedman's test [15], a non-parametric randomized block analysis of variance whose null hypothesis $H_0$ is that the error distributions are the same across repeated measures. [1] If the test rejects the $H_0$ hypothesis, a post-hoc analysis is run to find which pairs of methods are significantly different [36]. The analysis is based on Tukey's test and supplies an upper diagonal square matrix with the column element sorted by their rank.

Eventually, once the Friedman's test with post-hoc analysis has returned a rank of the different competing models, we may visualize the performance of all the models by means of boxplot/heat-map graphs.

## Case Study

This study applies the methodology discussed in Section 2 to a wind farm, located in southern Italy on the ridge of Apennines chain, composed of 15 wind generators and with an installed power of 30 MW. The forecasting concerns multiple time horizons (ranging from 1 to 6 hours).

[1] Note that such test differs from the conventional Analysis of variance (ANOVA) since it does not rely on any assumption of normal distribution and equal variances of residual.

**Table 3** Matrices $\mathbf{X_0}, \mathbf{Y_0}$, main features at changing of forecasting horizon

| forecasting horizon | | other parameters | | | | |
|---|---|---|---|---|---|---|
| hour | $H$ (number of step ahead) | L(H) Lag | $\phi$ | $\rho$ | $c$ | $\Phi(H)$ |
| 1 | 6 | 6 | 7 | 1 | 4 | 186 |
| 2 | 12 | 12 | 7 | 1 | 4 | 372 |
| 3 | 18 | 18 | 7 | 1 | 4 | 558 |
| 4 | 24 | 24 | 7 | 1 | 4 | 744 |
| 5 | 30 | 30 | 7 | 1 | 4 | 930 |
| 6 | 36 | 36 | 7 | 1 | 4 | 1116 |

The raw data includes wind speeds/directions at different heights, supplied by two anemometer spots, and the generated power, with a time resolution of 10 minutes over a period of 2 years. As shown in Fig. 2, the terrain is characterized by ridge steeps and complex orography, which causes chaotic behavior (notably fast changes in wind direction, strong shears, turbulence, sudden gusts). The impact of site morphology is confirmed by the spatial wind distribution shown by the wind roses of the two anemometers, where dominant winds come from south-west and north for anemometer 1 and 2, respectively. This setting increases the difficulties in relating anemometers measures with the wind farm power generation, requiring then the adoption of complex models to return an accurate forecasting.

Raw data have been processed according to the procedure of Section 2 then generating $V = 17$ experimental cases, whose features are shown in Table 3. Each test case is made of 25000 samples partitioned according a $5 : 1$ ratio into training and validation data-sets. Note that the training sets are processed by mRMR (Section 2) which returns the 15 most relevant variables.

## Experimental Results

The assessment results in terms of all the considered metrics and horizons are illustrated by a number of heat-maps (Figs. 3, 4 and 5) and boxplots (Figs. 4, 6 and 7). The heat-maps illustrate the ranking of WPF models according to the Friedman's test where the most accurate models are situated at the bottom-left side. Each cell of the heat-map takes two possible colors: grey if the WPF model in the row is significantly worse than the one in the column, orange otherwise. Such representation allows to visualize clusters of equivalent predictors (Figs. 3, 4 and 5).

To understand well the figures, it is necessary to comply to the following instructions: given a forecasting horizon

$H$, the reader should enter in the figure from downside by selecting a model and move from bottom to up until the diagonal cell, thus move perpendicularly from left to right over the row. All the orange elements are not significantly different models over this trip. For example, in case $H = 1\ h$ *GBM* lies in $6^{th}$ position in according to Friedman's test. Thus from bottom to the diagonal element there are two orange cells over the selected column (the diagonal element is always excluded from the count because it corresponds to the considered model itself), which means that the previous two models (*Av-GBM-RF* and *Ad-RF-Naive*) in the rank are not significantly different to *GBM*. Then, from the diagonal element, there are three orange cells by moving from left to right, which means that the successive three models in the rank (*Av-GBM-Naive*, *Av-SVM-GBM* and *RF*) are not significantly equal to *GBM*.

The information provided by the heat-maps, based on the accuracy ranks, are coupled with that provided by box-plots, which show the metric distribution for each considered WPF model and forecasting horizon. In this way, the decision maker will have a complete summary about the ranks and performance dispersion for each method leading to many interesting considerations made on the basis of the experimental results:

- the method which appears more consistently on the top ranking positions is the adaptive ensemble model combining RF, SVM, GBM and Naïve.
- Exponential Smoothing has the highest accuracy only for $H = 1$ h forecasting horizon whereas for increasing $H$ its performances deteriorates dramatically in terms of MSE, which tends to penalize greater the large prediction error respect with MAE.
- the ensemble forecasting of ANNs plus Naïve tends to have higher ranking at increasing of forecasting horizon $H$.

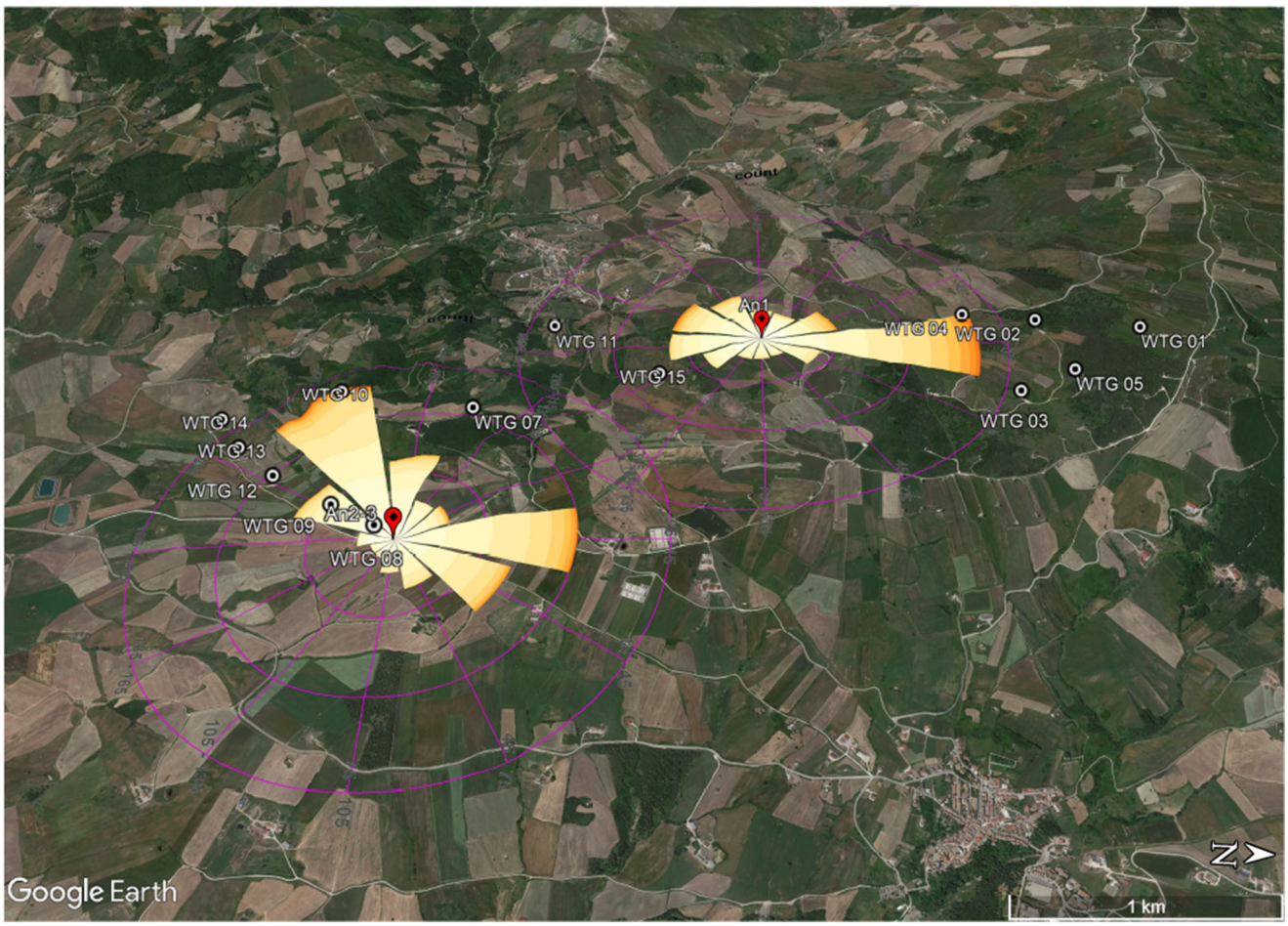


**Fig. 2** Wind farm satellite view

– The influence of the system physics becomes dominant in $H = 5$ and $H = 6$, causing a general reduction of accuracy in all the WPF models. Indeed, the spreads become much wider and there is no more a single model outperforming the rest.

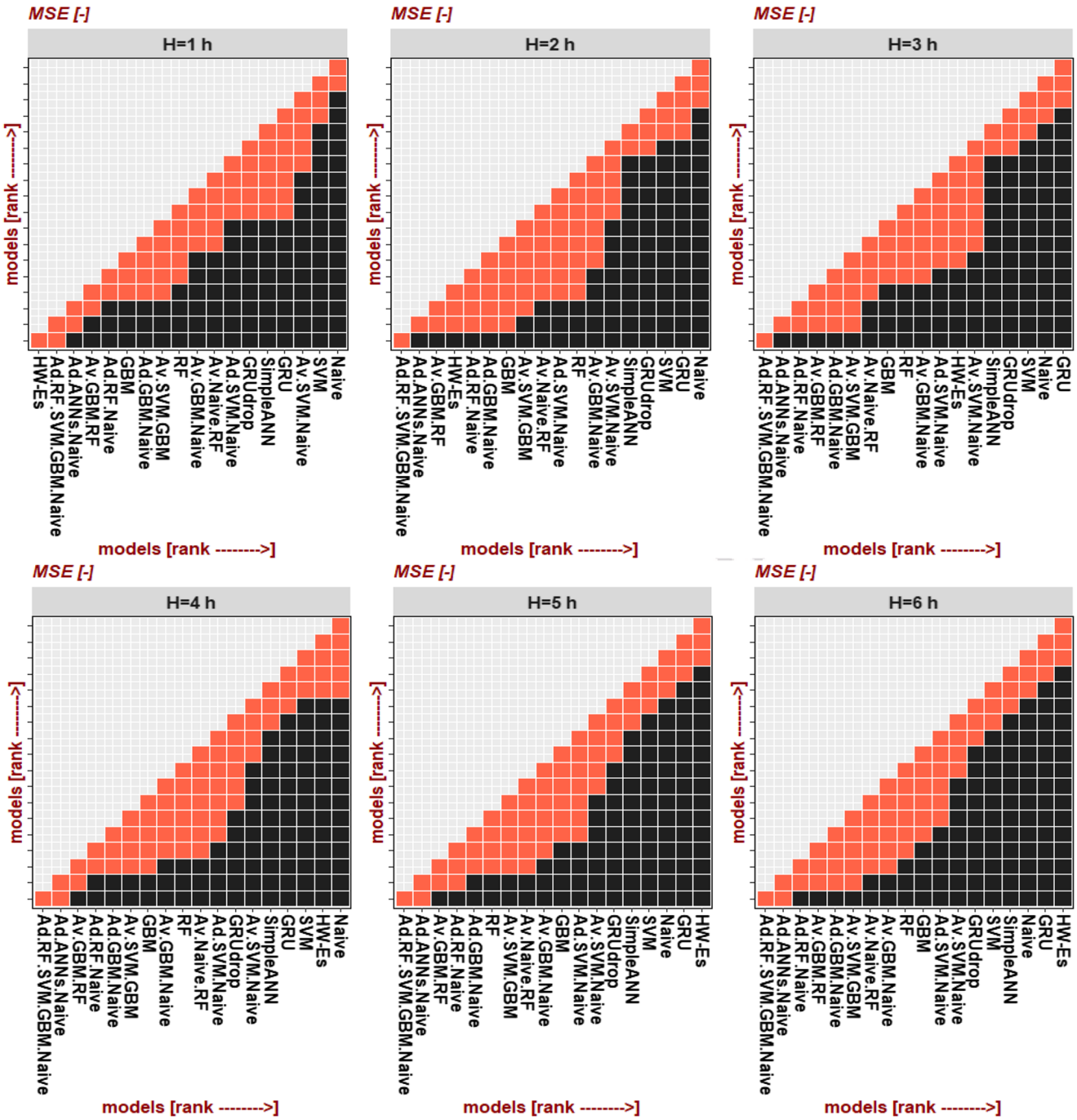


**Fig. 3** Visualization of Friedman's test with Post-hoc Analysis in terms of MSE. The best models are on the leftmost/lower side. The elements placed on the left side are placed in a specular manner to those on the downside for each insert. An orange case means that the models on the respective row/column are not significantly different

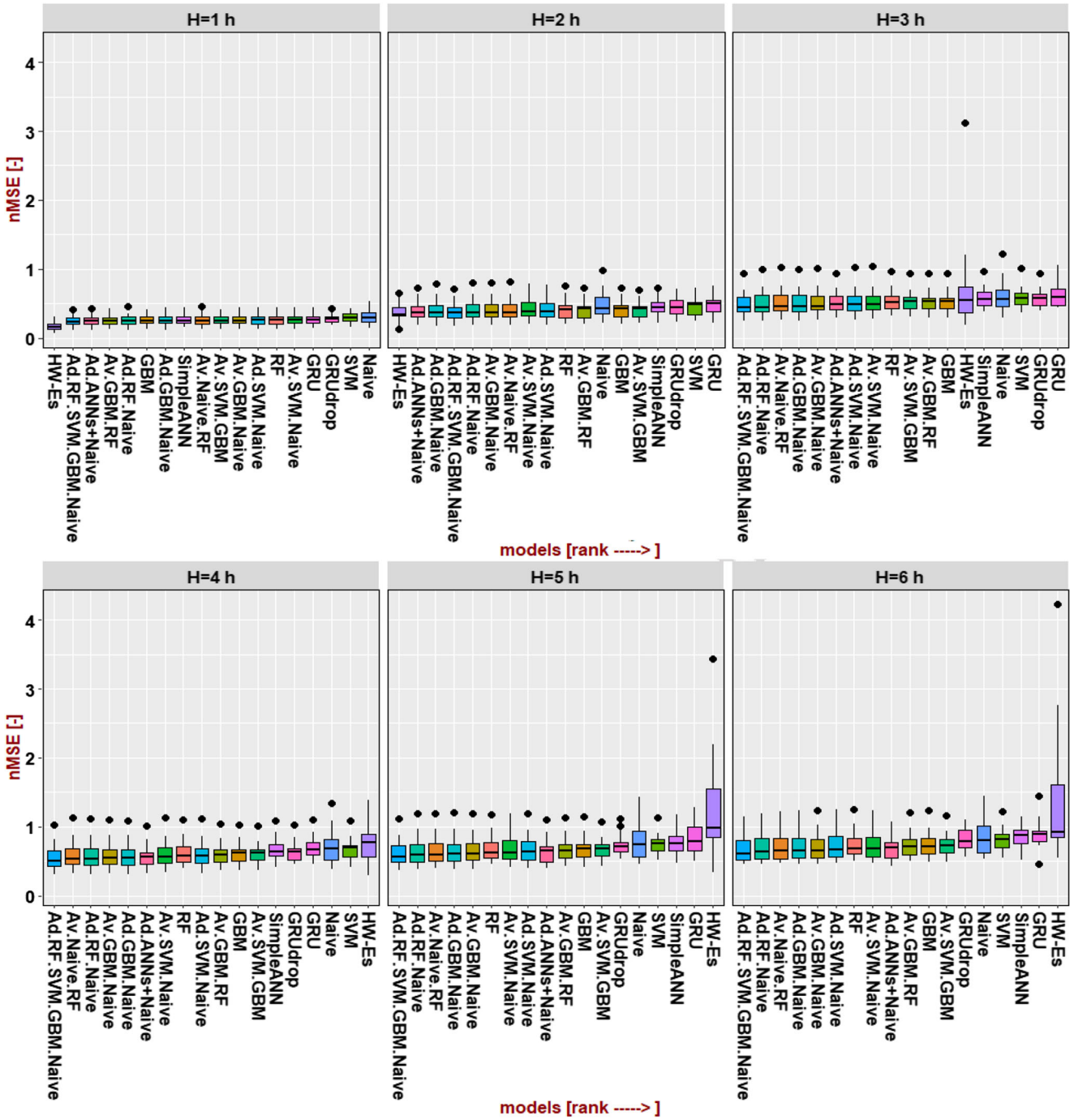


**Fig. 4** Visualization of model performance distribution in terms of nMSE. The models are ordered by considering the rank supplied by the Friedman's test

On the basis of the previous considerations, our recommendation would be the adoption of ensemble forecasting, which is able to guarantee both accuracy and stability over different forecasting settings.

## On-going Experimental Activities

The proposed approach has been developed to predict the overall wind farm power generation. Anyway, in the

preliminary design phase, also a wind power forecasting on the single wind generator has been tested, with worse results. This phase had the merit to highlighting the need for enhancing the condition monitoring of the generator asset, improving the level of confidence in the analysis of the operation data. In particular, the comparison between the output of the wind turbine generator models and the measured data allowed detecting several critical operation

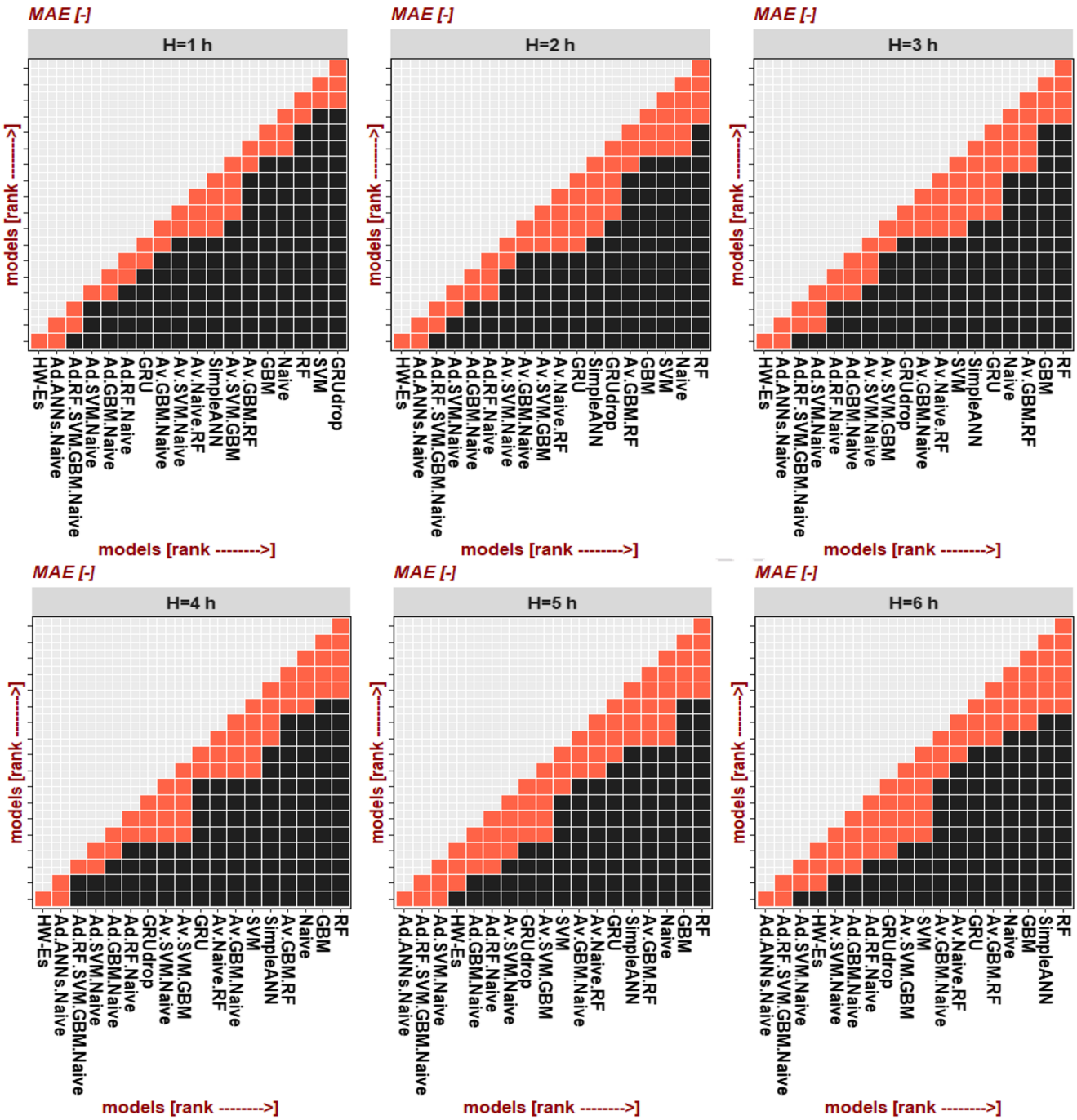


**Fig. 5** Visualization of Friedman's test with Post-hoc Analysis in terms of MAE. The best models are on the leftmost/lower side. The elements placed on the left side are placed in a specular manner to those on the downside for each insert. An orange case means that the models on the respective row/column are not significantly different

states, especially in analyzing the wind - power output with a 10-min time resolution. Indeed, as shown by Fig. 8, the active power spread for each wind speed bin is significant, especially for low values. This is caused by the effect of generator inertia, which smooths the fast transients in wind gusts as shown in the lower side of the figure. Actually, the overall combined effect of inertia, blade/wind peripheral rotor speed and yaw controls is the main factor inducing this power spreading [47]. Further, critical patterns include the deflections from the maximum

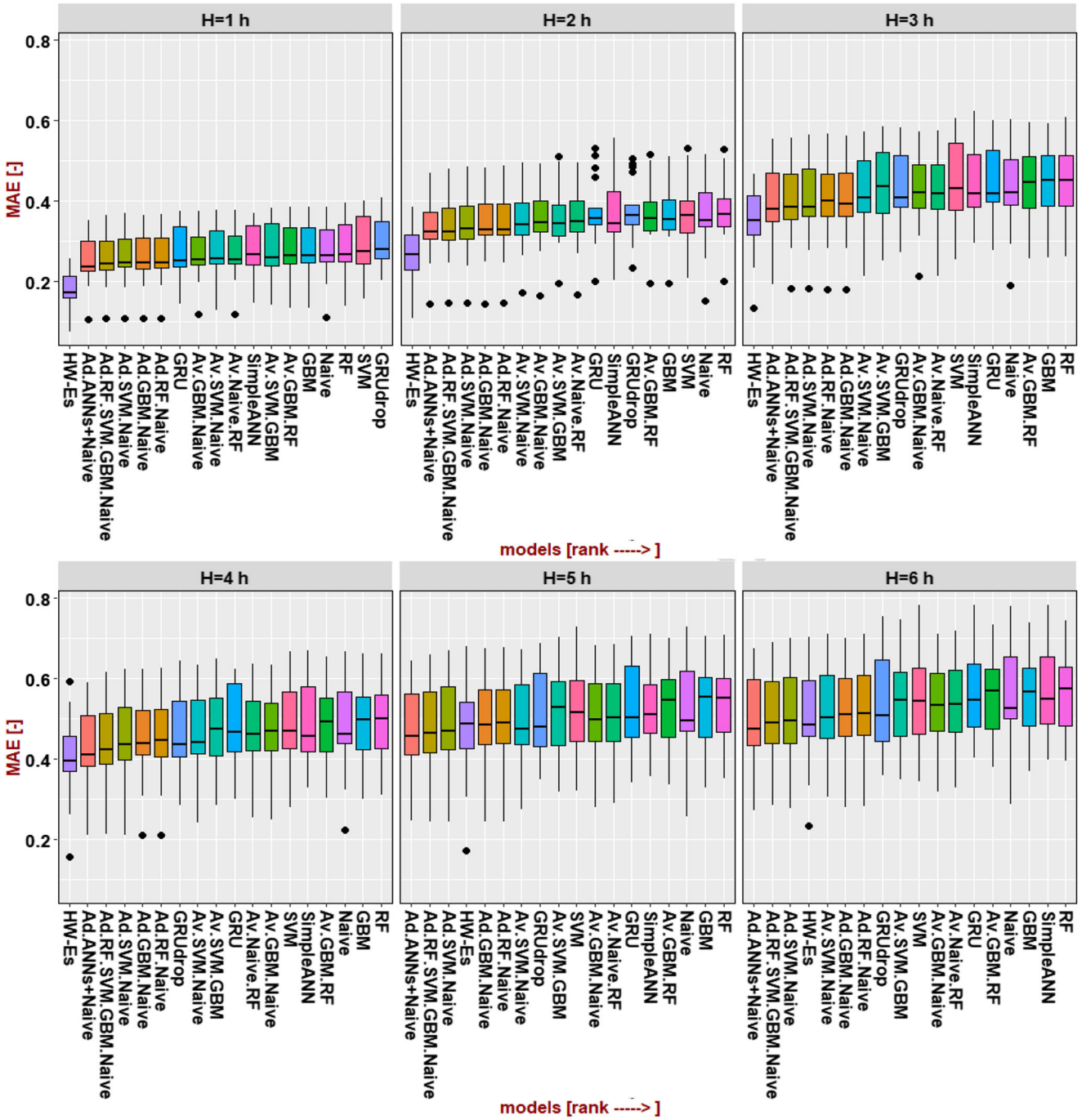


**Fig. 6** Visualization of model performance distribution in terms of MAE. The models are ordered by considering the rank supplied by the Friedman's test

power operation point for the wind speeds greater than the nominal value (rightmost side), where several data clusters can be clearly identified. These derated states are activated in order to satisfy the power curtailment orders imposed by the transmission system operator in the presence of power system congestion. Any other deviation between the predicted and the measured data is a clear indication of a system anomaly, which could be induced by wrong control settings, performance deterioration, incipient faults etc. A data-driven methodology aimed at classifying

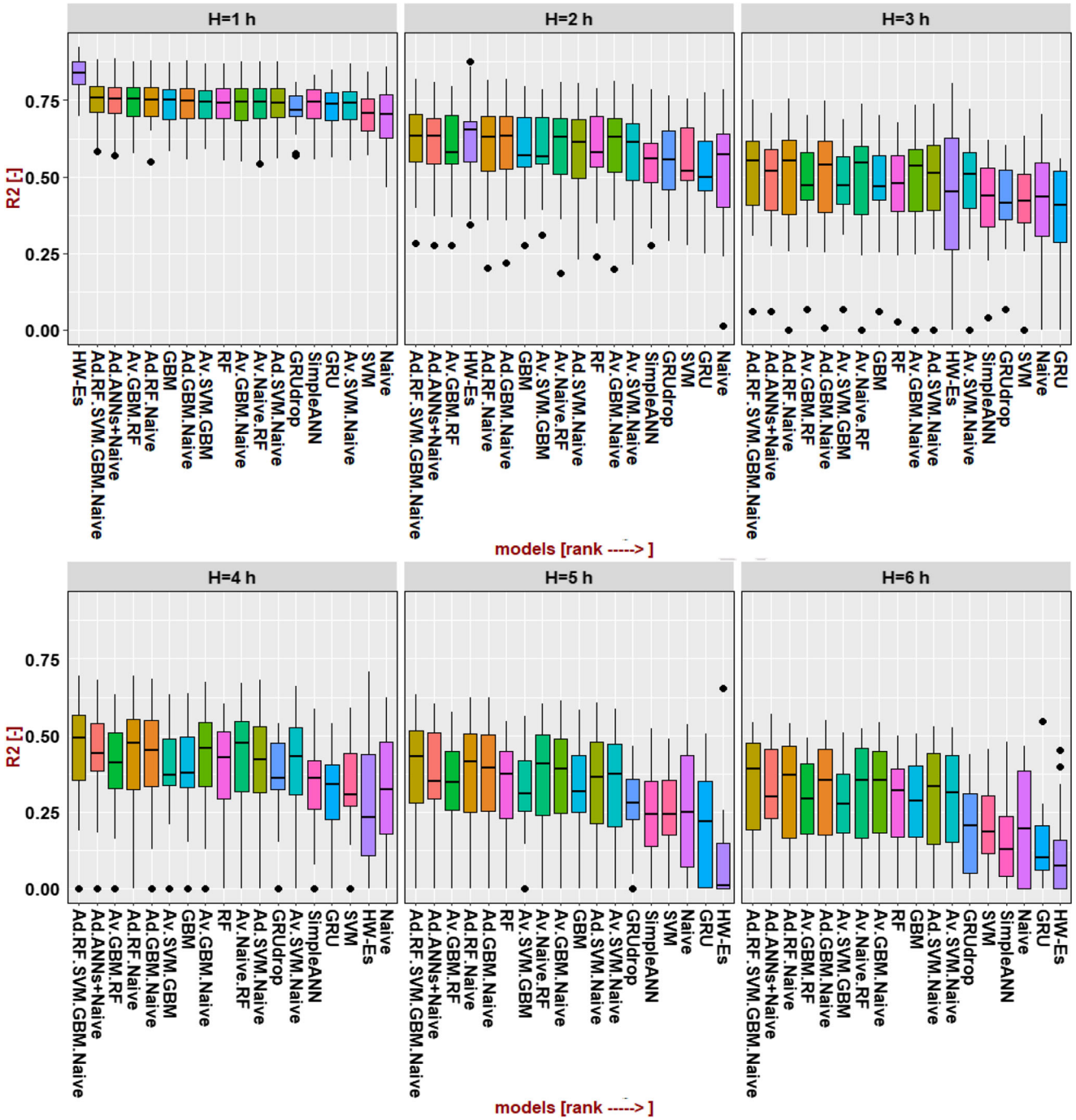


**Fig. 7** Visualization of model performance distribution in terms of $R^2$. The models are ordered by considering the rank supplied by the Friedman's test

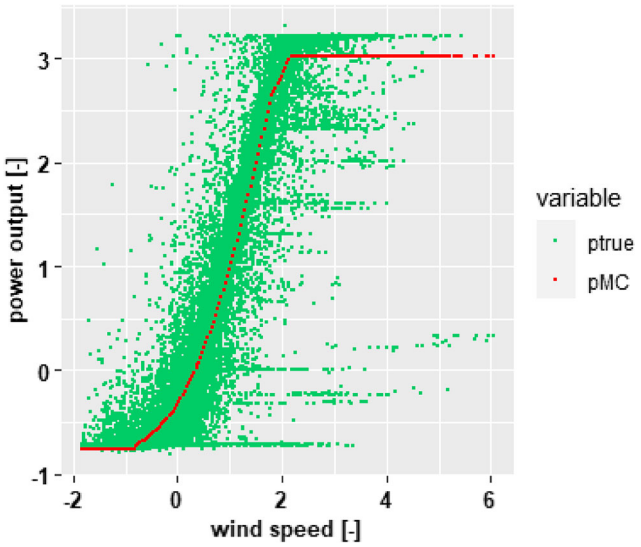


**Fig. 8** Comparison between WTG experimental measures ("ptrue") and manufacturer curve ("pMC")

these events are currently under development by the Authors.

## Conclusions

The growing wind power production is introducing uncertainty in power grid operations with negative consequences for both system operators and wind producers because of the stochastic and intermittent nature of the wind. This issue motivates the design of effective wind power forecasting models to mitigate the power generation uncertainty. Nevertheless a thorough procedure for comparing and assessing existing approaches is still lacking in literature on wind forecasting.

This paper proposed a methodology for the robust assessment of different types of short-term WPF models over multiple horizons, which is based on data resampling and statistical tests.

The study highlighted that ensemble forecasting of statistical and machine learning models dominates in terms of accuracy ranks, by supplying additional robustness with respect to single approaches.

In particular, for horizons longer than two hours, such technique is able to outperform exponential smoothing a well-known state-of-the-art statistical approach.

As far as the generalization of these results is concerned, it is important to remark that the pipeline and the features selection technique have been designed to process heterogeneous data sets, characterized by different size and complexity. Moreover, additional time series, i.e. temperature and pressure profiles, can be integrated in the input data-set and processed by the forecasting algorithms, if their contribution is considered relevant by the feature selection technique.

Future research will focus on the improvement the performance of ensemble forecasting models. This kind of models will be adapted to supply wind power forecast on large area, trying to catch correlation between the several wind power plants by supplying multivariate and hierarchical forecast over the time.

**Acknowledgements** This research was supported by OSMOSE (Optimal System-Mix Of flexibility Solutions for European electricity) project, in relation to Horizon 2020 research and innovation programme under grant agreement No 773406.

## Compliance with Ethical Standards

**Conflict of interests** The authors declare that they have no conflict of interest.

## References

1. Albadi M, El-Saadany E (2010) Overview of wind power intermittency impacts on power systems. Electric Power Systems Research 80(6):627–632. http://www.sciencedirect.com/science/article/pii/S0378779609002764
2. Amjady N, Keynia F, Zareipour H (2011) Wind power prediction by a new forecast engine composed of modified hybrid neural network and enhanced particle swarm optimization. IEEE Transactions on Sustainable Energy 2(3):265–276
3. Balcilar M, Ozdemir ZA, Arslanturk Y (2010) Economic growth and energy consumption causal nexus viewed through a bootstrap rolling window. Energy Economics 32(6):1398–1410. http://www.sciencedirect.com/science/article/pii/S0140988310000952
4. Bontempi G, Taieb SB (2011) Conditionally dependent strategies for multiple-step-ahead prediction in local learning. International Journal of Forecasting 27(3):689–699
5. Bontempi G, Taieb SB, Le Borgne YA (2012) Machine learning strategies for time series forecasting. In: European business intelligence summer school. Springer, pp 62–77
6. Bottou L, Vapnik V (1992) Local learning algorithms. Neural Comput 4(6):888–900. https://doi.org/10.1162/neco.1992.4.6.888
7. Cameron AC, Windmeijer FA (1997) An r-squared measure of goodness of fit for some common nonlinear regression models. J Econometrics 77(2):329–342
8. Cardell J, Anderson L, Tee CY (2010) The effect of wind and demand uncertainty on electricity prices and system performance. In: IEEE PES T D 2010, pp 1–4
9. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555
10. De Stefani J, Le Borgne YA, Caelen O, Hattab D, Bontempi G (2019) Batch and incremental dynamic factor machine learning for multivariate and multi-step-ahead forecasting. Int J Data Sci Analytics 7(4):311–329
11. Demolli H, Dokuz AS, Ecemis A, Gokcek M (2019) Wind power forecasting based on daily wind speed data using machine learning algorithms. Energy Conversion and Management 198:111823
12. Ela E, O'Malley M (2012) Studying the variability and uncertainty impacts of variable generation at multiple timescales. IEEE Trans Power Sys 27(3):1324–1333

13. Foley AM, Leahy PG, Marvuglia A, McKeogh EJ (2012) Current methods and advances in forecasting of wind power generation. Renewable Energy 37(1):1–8
14. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of Statistics, pp 1189–1232
15. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701. https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522
16. Ghiasi M, Esmaeilnamazi S, Ghiasi R, Fathi M (2020) Role of renewable energy sources in evaluating technical and economic efficiency of power quality. Technology and Economics of Smart Grids and Sustainable Energy 5(1):1
17. Gielen D, Boshell F, Saygin D, Bazilian MD, Wagner N, Gorini R (2019) The role of renewable energy in the global energy transformation. Energy Strategy Reviews 24:38–50. http://www.sciencedirect.com/science/article/pii/S2211467X19300082
18. González-Aparicio I, Monforti F, Volker P, Zucker A, Careri F, Huld T, Badger J (2017) Simulating european wind power generation applying statistical downscaling to reanalysis data. Appl Energy 199:155–168
19. Gooijer JGD, Hyndman RJ (2006) 25 years of time series forecasting. Int J Forecasting 22(3):443–473. http://www.sciencedirect.com/science/article/pii/S0169207006000021, twenty five years of forecasting
20. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
21. Hao Y, Tian C (2019) A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. Appl Energy 238:368–383
22. Holt CC (2004) Forecasting seasonals and trends by exponentially weighted moving averages. Int J Forecasting 20(1):5–10. http://www.sciencedirect.com/science/article/pii/S0169207003001134
23. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecasting 22(4):679–688. http://www.sciencedirect.com/science/article/pii/S0169207006000239
24. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37
25. Kavasseri RG, Seetharaman K (2009) Day-ahead wind speed forecasting using f-arima models. Renewable Energy 34(5):1388–1393. http://www.sciencedirect.com/science/article/pii/S0960148108003327
26. Korprasertsak N, Leephakpreeda T (2019) Robust short-term prediction of wind power generation under uncertainty via statistical interpretation of multiple forecasting models. Energy 180:387–397
27. Lee S, Lee DK (2018) What is the proper way to apply the multiple comparison test? Korean Journal of Anesthesiology 71(5):353
28. Lian X, Chen L (2009) General cost models for evaluating dimensionality reduction in high-dimensional spaces. IEEE Trans Knowl Data Eng 21(10):1447–1460
29. Liang Z, Liang J, Wang C, Dong X, Miao X (2016) Short-term wind power combined forecasting based on error forecast correction. Energy Conversion and Management 119:215–226
30. Liaw A, Wiener M et al (2002) Classification and regression by randomforest. R news 2(3):18–22
31. Liu Y, Guan L, Hou C, Han H, Liu Z, Sun Y, Zheng M (2019) Wind power short-term prediction based on lstm and discrete wavelet transform. Appl Sci 9(6):1108
32. Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: Concerns and ways forward, vol 13
33. Mararakanye N, Bekker B (2019) Renewable energy integration impacts within the context of generator type, penetration level and grid characteristics. Renewable and Sustainable Energy Reviews 108:441–451
34. Mendes-Moreira J, Soares C, Jorge AM, Sousa JFD (2012) Ensemble approaches for regression: A survey. Acm Computing Surveys (csur) 45(1):10
35. Ozkan MB, Karagoz P (2015) A novel wind power forecast model: Statistical hybrid wind power forecast technique (shwip). IEEE Trans Industrial Informatics 11(2):375–387
36. Pereira DG, Afonso A, Medeiros FM (2015) Overview of friedman's test and post-hoc analysis. Communications in Statistics-Simulation and Computation 44(10):2636–2653
37. Ren Y, Suganthan P, Srikanth N (2015) Ensemble methods for wind and solar power forecasting—a state-of-the-art review. Renewable and Sustainable Energy Reviews 50:82–91. http://www.sciencedirect.com/science/article/pii/S1364032115003512
38. Rong M, Gong D, Gao X (2019) Feature selection and its use in big data: Challenges, methods, and trends. IEEE Access 7:19709–19725
39. Soares T, Pinson P, Jensen TV, Morais H (2016) Optimal offering strategies for wind power in energy and primary reserve markets. IEEE Transactions on Sustainable Energy 7(3):1036–1045
40. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
41. Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. Int J Forecasting 16(4):437–450
42. Torres J, García A, Blas MD, Francisco AD (2005) Forecast of hourly average wind speed with arma models in Navarre (Spain). Solar Energy 79(1):65–77. http://www.sciencedirect.com/science/article/pii/S0038092X04002877
43. Wan C, Xu Z, Pinson P, Dong ZY, Wong KP (2013) Optimal prediction intervals of wind power generation. IEEE Transactions on Power Systems 29(3):1166–1174
44. Wang HZ, Li GQ, Wang GB, Peng JC, Jiang H, Liu Y (2017) Deep learning based ensemble approach for probabilistic wind power forecasting. Applied Energy 188:56–70
45. Würth I, Valldecabres L, Simon E, Möhrlen C, Uzunoğlu B, Gilbert C, Giebel G, Schlipf D, Kaifel A (2019) Minute-scale forecasting of wind power—results from the collaborative workshop of iea wind task 32 and 36. Energies 12(4):712
46. Xu Q, He D, Zhang N, Kang C, Xia Q, Bai J, Huang J (2015) A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining. IEEE Transactions on Sustainable Energy 6(4):1283–1291
47. Yan J, Liu Y, Han S, Wang Y, Feng S (2015) Reviews on uncertainty analysis of wind power forecasting. Renewable and Sustainable Energy Reviews 52:1322–1330
48. Zeng J, Qiao W (2011) Support vector machine-based short-term wind power forecasting. In: 2011 IEEE/PES Power Systems Conference and Exposition. IEEE, pp 1–8
49. Zhang Y, Wang J, Wang X (2014) Review on probabilistic forecasting of wind power generation. Renewable and Sustainable Energy Reviews 32:255–270. http://www.sciencedirect.com/science/article/pii/S1364032114000446