DAFT-E: Feature-based Multivariate and Multi-step-ahead Wind Power Forecasting

Article *in* IEEE Transactions on Sustainable Energy · November 2021

DOI: 10.1109/TSTE.2021.3130949

4 authors:

Fabrizio De Caro
Università degli Studi di Salerno
54 PUBLICATIONS **169** CITATIONS

Jacopo De Stefani
Université Libre de Bruxelles
15 PUBLICATIONS **111** CITATIONS

Alfredo Vaccaro
Università degli Studi del Sannio
235 PUBLICATIONS **4,302** CITATIONS

Gianluca Bontempi
Université Libre de Bruxelles
376 PUBLICATIONS **17,805** CITATIONS

DAFT-E: Feature-based Multivariate and Multi-step-ahead Wind Power Forecasting

Fabrizio De Caro *†, *Member, IEEE,* Jacopo De Stefani *‡,
Alfredo Vaccaro †, *Senior Member, IEEE,* and Gianluca Bontempi ‡

***Abstract*—Wind energy is one of the most promising resources for the mitigation of greenhouse gas emissions that contribute to anthropogenic global warming. However, the large proliferation of wind power generators is causing several critical issues in power systems due to their variable power generated profiles. For this reason, a large number of learning techniques, e.g. integrating Vector Auto-Regressive and Neural Network-based models, were proposed in the literature for mitigating wind power uncertainty issues. Unfortunately, these methodologies show several limitations, e.g. the huge number of parameters and/or the heavy computational cost, which hinder their deployment in modern power system operation, where prompt and reliable wide-area wind power generation forecasts are requested for supporting time-critical decision making on several time horizons. To try addressing this issue, this paper proposes the Dynamic Adaptive Feature-based Temporal Ensemble (DAFT-E) forecasting approach, which relies on an extensive feature engineering, a fast feature selection step and an ensemble of computationally inexpensive models to reduce the computational complexity of the forecasting task, while still preserving predictive accuracy. The experimental results, which benchmark DAFT-E against multivariate (VAR and deep learning) alternatives on two real case studies, show that the proposed approach outperforms state-of-the-art and representation learning models according to several forecasting accuracy metrics.**

***Index Terms*—Wind Power Forecasting, Power System Operations, Machine Learning, Spatio Temporal Features, Forecasting Model Validation, Ensemble Forecasting.**

I. Introduction

WIND power generation is one of the most promising energy sources to reduce the impact of power systems on global warming. Recently, the wind power technological competitiveness and the incentive policies implemented by authorities allowed a proliferation of Wind Power Generators (WPG) in power systems [1].

Unfortunately, WPGs introduce uncertainty in power systems due to their variable behavior. The uncertainty components are linked to wind behavior and to wind energy generation [2]. First, although wind dynamics are well understood, modeled, and predicted, the wind magnitude and direction exhibit random fluctuations according to the considered time scale [3]. Second, operative conditions, mechanical aging, aerodynamic interference among wind power generators, and rotor inertia affect the power generation output given a set of weather predictions [4]. These uncertainty sources push the system operators to revise their traditional planning and operation tools to mitigate the impacts induced by a massive grid integration of WPGs. In this scenario, accurate wind power forecasting represents an effective tool for managing the generated power fluctuations, improve resources provisioning and ensure safe system operations [5]. Indeed, the transmission system operator (TSO) needs spatial and temporal forecasting of load and renewable energy sources (RES) over a time window ranging from 1-week to 5-min. In short-term system operation (pre-dispatch phase), which extends from 1-week to a day ahead of actual operation, TSO needs predictions to identify and allocate the system reserve. In real-time operation, which extends from 5-min to 30-min ahead of actual operation, TSO needs predictions to bear RES/load increments/decrements following deviations from the predicted profiles; to evaluate the power system status; to take preventive actions for assuring a secure and reliable grid operation [6].

Since these decision processes concern events typically occurring over a horizon ranging from 5 minutes to 6 hours, accurate short-term wind power forecasting on large areas is strategic for anticipating critical events [7].

The literature proposed several forecasting methodologies, where the dynamic model adaptation represents one of the most important issues to address in order to model the intrinsic time-varying behaviour characterizing the wind dynamics [5].

Wide area forecasting is often approached as a multivariate forecasting problem, where each wind farm generation profile represents a variable. The most adopted approaches are statistical and Machine Learning (ML) based. Statistical approaches include vector regressions (VAR, VARMA, VARIMA, VARMAX) [8], as well as kernel-based regression [9]. In machine learning we can distinguish between feature-engineering techniques, requiring expert know-how to encode useful information into input features, and representation-based techniques, where the model is expected to discover the optimal data representation during the learning process [10].

Vector AutoRegressive (VAR) models showed a good capability in capturing linear dependencies between wind

(†) F. De Caro and A. Vaccaro are affiliated with the Department of Engineering, University of Sannio, Piazza Roma 21, 82100, Benevento, Italy e-mail: [fdecaro,vaccaro]@unisannio.it

(‡) J. De Stefani and G. Bontempi are affiliated with the Machine Learning Group (MLG), Department of Computer Science, Universitè Libre de Bruxelles, Campus de la Plaine ULB CP212, boulevard du Triomphe, 1050 Bruxelles, Belgium, [jacopo.de.stefani,gbonte]@ulb.ac.be

(*) Fabrizio De Caro and Jacopo De Stefani are co-first authors (Corresponding author: Fabrizio De Caro).

Digital Object Identifier 10.1109/TSTE.2021.3130949,©2021 IEEE

farms [11]. Since canonical VAR models consider stationary power dynamics, [12] proposed an adaptive lasso VAR based model with Forgetting Factors to improve the prediction accuracy. The authors of [13] proposed a correlation constrained and sparsity controlled VAR to reduce the effective number of parameters in model training. Indeed, the main VAR-based model drawback is the dramatic parameter growth at the increasing of the time series and lag sample number according to $(N+L\cdot N^2)$, where N and L are the number of time series and lags, respectively.

In the machine learning community, representation based Deep-Learning (DL) is more and more used in wind and power forecasting [14] because of its success with non-linear spatial relationships [15]. In particular, Recurrent and Convolutional Neural Networks are the most promising models for predicting wind power time series [16]. Unfortunately, their lack of robustness in highly dynamic settings, as well as the need for specialized architecture for efficient computational performance and extensive fine-tuning [17], makes them unsuitable for time-critical applications, notably short-term forecasting of many wind farms for system monitoring purposes. Moreover, the lack of interpretability of the model and the automatically determined features hinder the extraction of useful information for forward planning applications.

At the same time, the recent advances in feature selection (also for very large dimension settings) suggest that a feature engineering strategy should be considered as an interesting alternative to black box solutions, mainly if interpretability, robustness and computational time are at stake.

To the best of our knowledge, this paper is the first systematic comparison of feature engineering and representation learning strategies for multivariate wind power forecasting.

In particular, the main contributions of this manuscript are:

1) The design of a novel ensemble forecasting methodology based on model aggregation and feature engineering, adaptive error-based combination weights and Forgetting Factors (FFs);
2) The introduction of novel features able to detect specific operating conditions as wind power generation curtailment or null production, which can deteriorate the model training in a simulated real condition environment;
3) A model assessment procedure based on the bias-variance principle, which highlights the multivariate model performance over the space through a bivariate box plot visualization [18];
4) The comparison between different families of multivariate forecasting strategies over different horizons, time resolutions and taking into account, besides accuracy, financial risk measures as Value at Risk (VaR) and conditional Value at Risk (cVaR).

II. Multivariate Forecasting Pipeline

The modeling of a N-variate time series, where $y_{n,t}$ is the generic exchanged power generation profiles at MV/HV substations, requires a number of steps to obtain a reliable multi-step-ahead forecasting. Given a time resolution $\Delta t = t_i - t_{i-1}$, a time instant t, and a forecasting horizon span $\mathcal{H} = \{1, \dots, h, \dots, H\}$, where h and H are the generic and maximum forecasting horizon, a multi-variate and multi-temporal model f aims to estimate the expected future values conditional on the past behavior:

$$\begin{array}{l} y_{1,t+1}, \dots, y_{1,t+H} \\ \dots \\ y_{N,t+1}, \dots, y_{N,t+H} \end{array} = f\left(\begin{array}{c} y_{1,t-d-L}, \dots, y_{1,t-d} \\ \dots \\ y_{N,t-d-L}, \dots, y_{N,t-d} \end{array} \right) \quad (1)$$

where L, d, and N are the lag, delay, and the time series number, respectively. The multi-input multi-output (MIMO) nature of the mapping (1) can be addressed in several manners according to the assumptions made about the nature of the temporal and cross-series dependencies.

In the global approach, a single multi-input multi-output model is learned from the observed data. In the local approach, the forecasting problem is decomposed in several sub-tasks, which are addressed independently [19]. Note that the decomposition may occur at several levels: for instance we could decompose the MIMO mapping in N multi-input single-output (MISO) tasks or in N single-input single-output (SISO) tasks.

The nature of the multivariate problem implies that the best way to introduce a forecasting method is to present it as a computational pipeline addressing the design issues in a sequence of steps:

1) Pre-Processing: this step typically normalizes the observations and rescales them to a suitable temporal resolution according the considered context. Missing data may be replaced by applying spatial averaging techniques [20].
2) Feature Engineering: this step augments the representation space by constructing a number of additional input features capturing either the temporal dynamic of the signal or the occurrence of specific events (e.g. curtailment).
3) Embedding strategy: a forecasting problem may be set as a supervised input-output problem where the nature of the output and the dimension of the input space depend on the horizon H, temporal lag L and the cross-series dependencies taken into account [21].
4) Dimensionality reduction: this step aims to reduce the the number of features, by compression (via Principal Components Analysis - PCA) or by feature selection. This process reduces the risk of curse of dimensionality [22] and the computational burden in model training [23], by removing features that are redundant with respect to the other input variables or low correlated to the predicted variable.
5) Model estimation: this step estimates from the available data the input-output relationship defined in the previous steps. State-of-the-art approaches are discussed in Section IV-C.
6) Performance Assessment: this step typically splits the observed data into two parts: one for learning the input-output mapping (training set) and the other one to validate the model performance (validation set). The distribution of performance measures on the validation set is then analyzed in order to assess the correctness and

robustness of the forecasts both in typical and worst-case situations.

III. Proposed Methodology

This section introduces the proposed multi-input single-output (MISO) methodology by detailing the most peculiar steps of its computational pipeline.

1) Feature Engineering: first, this step augments the input space by computing a number of conventional statistics across a time window of the past w values:

$$\text{Moving Average} \quad \bar{y}_t = \frac{1}{w+1}\sum_{q=0}^{w} y_{t-q} \tag{2}$$

$$\text{Maximum Value} \quad y_t^{+} = \max_{q\in\{0,\cdots,w\}} y_{t-q} \tag{3}$$

$$\text{Minimum Value} \quad y_t^{-} = \min_{q\in\{0,\cdots,w\}} y_{t-q} \tag{4}$$

$$\text{p-quantile} \quad \begin{array}{c} y_{p,t} = \inf\{z : \widehat{F}_w(z) \geq p\} \\ z \in \{y_{t-0},\cdots,y_{t-q}\} \\ \widehat{F}_w(z) = \frac{1}{w}\sum_{q=0}^{w} \mathbf{1}_{y_{t-q}\leq z} \end{array} \tag{5}$$

$$\text{1st order difference} \quad \begin{array}{c} \Delta y_t = y_t - y_{t-1} \\ y_t \in \{y_{t-0},\cdots,y_{t-q}\} \end{array} \tag{6}$$

Second, it introduces parametric, expert-based features to detect the curtailment of wind power series, which causes the deflection of the trajectories from a certain dynamical trend to a constant one. These features are constructed, respectively, using first order difference based method, discarding all signal variability smaller than σ (7), and a Run Length Encoding (RLE) based detection, employing the auxiliary indicator function $\mathbf{1}^S(\cdot)$ (8), discarding all the sequences of constant values shorter than a given parameter v (9). Both the parameters σ and v are externally specified.

$$\mathbf{1}_{\sigma}^{FOD}(y_t) = \begin{cases} 1 & |\Delta y_t| < \sigma \\ 0 & otherwise \end{cases} \tag{7}$$

$$\mathbf{1}^{S}(y_t) = \begin{cases} 1 & \Delta y_t = 0 \\ 0 & otherwise \end{cases} \tag{8}$$

$$\mathbf{1}_{v}^{RLE}(y_t) = \begin{cases} 1 & \exists t_s > t_0, t_s < t_e < t \text{ s.t.} \\ & t_e - t_s > v \wedge \forall i \in \{t_s \cdots, t_e\} \mathbf{1}^S(y_i) = 1 \\ 0 & otherwise \end{cases} \tag{9}$$

These derived features augment the cardinality of $\mathbf{Y}[S\times N]$, where S is the available number of samples and N is the number of input time series. After that, the augmented number of time series is $N' = N\cdot(1+n_s\cdot s_q+n_f)$, where n_s is the number of statistics (2)-(5) computed for s_q different lags, and n_f is the number of features (6)-(9).

2) Data Embedding: The data embedding process rearranges $\mathbf{Y}$ in the matrices of targets $\mathbf{R}$ and predictors $\mathbf{P}$, as described in [24], given the parameters L, d, and H. In particular, the obtained matrices are $\mathbf{P}[S'\times N']$ and $\mathbf{R}[S'\times F]$, where $S' = S-(L+d+H+1)$ and $F = N\cdot H$. Hence, the latter matrices are split into training and test matrices producing $\mathbf{R}_{trn}[S_{trn}\times F]$, $\mathbf{R}_{test}[S_{test}\times F]$, $\mathbf{P}_{trn}[S_{trn}\times N']$, and $\mathbf{P}_{test}[S_{test}\times N']$, where $S_{test} = S' - S_{trn}$.

3) Feature selection: The large dimensional space, due to the combined effect of smoothing and embedding processes, may be addressed by filter selection techniques like the minimum Redundancy Maximum Relevance (mRMR) [25]. mRMR returns a subset of $N_S << N'$ relevant features by using a forward procedure which at the g-th step ($g = 1,\ldots,N_S$) selects the least redundant and most informative predictor variable:

$$\operatorname*{arg\,max}_{\mathbf{P}_\lambda\in P-\Phi_{g-1}} \left[I(\mathbf{P}_\lambda;\mathbf{r}) - \frac{1}{g-1}\sum_{\mathbf{P}_\psi\in\Phi_{g-1}} I(\mathbf{P}_\psi;\mathbf{P}_\lambda) \right] \tag{10}$$

where Φ_{g-1} is the set of $g-1$ previously selected variables and $I(x,y)$ denotes the mutual information [26]. Note that the mutual information term can be efficiently estimated by $I(x,y) = 1/2\ln(1-\rho(x,y)^2)$ where ρ is the Pearson correlation coefficient under an assumption of normality. A specific advantage of mRMR with respect to compression techniques is that it does not transform the original features, allowing an easier data interpretation.

4) Dynamic Adaptive Feature-based Temporal Ensemble: We propose an original method, called Dynamic Adaptive Feature-based Temporal Ensemble (DAFT-E), based on the weighted average of M forecasting models, whose weights evolution depends on their forecasting errors over a sliding window of size Γ and a forgetting strategy. The pseudo-code of the method is detailed in the Algorithm 1.

The multivariate multi-step-ahead problem is decomposed in a set of $F = N\cdot H$ multi-input single-output tasks by applying a direct strategy [21] for each n-th column of $\mathbf{R}$. Each prediction task $f_{n,h}$ (line 6) is addressed by M algorithms and the M predictions are combined by weighted averaging (line 18).

Every Γ steps, the weights are returned by the inverse of the mean of the latest Γ squared forecasting errors (line 10), then normalized (line 15) and eventually regularized (line 16).

The regularization uses a vector of V forgetting factors (FFs) $\Lambda_1,\ldots,\Lambda_V$ satisfying $0 < \Lambda_v < 1, \sum_{v=1}^{V}\Lambda_v = 1$, which quantifies the contribution of the V previous cycles.

DAFT-E controls the bias/variance trade-off of the adaptive algorithm thanks to the hyperparameters Γ and $\Lambda_v, v = 1,\ldots,V$. The larger Γ and the more similar the Λ_v values, the higher the smoothness (and consequently the bias) of the forecast estimation. Such dynamic regularization process allows better robustness (and then accuracy) in front of cyclic regime changes (ramps, power generation curtailments). Given the high degree of uncertainty of the wind process, the memory-based weights update process reduces the sensitivity to recent noise values and decreases variance and instability. In

Algorithm 1 DAFT-E algorithm for the k-th trial having V weight update cycles with FF vector $\mathbf{\Lambda}$ and window size Γ

```
    Input: $M$ Algorithms, $\mathbf{P}^{(k)}_{trn}, \mathbf{R}^{(k)}_{trn}, \mathbf{P}^{(k)}_{test}, \mathbf{\Lambda}, \Gamma$
    Output: $\hat{\mathbf{X}}$ (DAFT-E prediction)

$N_c \leftarrow \lfloor N_{test}/\Gamma \rfloor$
▷ Update the weights over the time span $\Gamma$
for $c \leftarrow V$ to $N_c$ do
▷ when $V < c$ the $\mathbf{w}_{norm}$ are initialized
    $t_{start} \leftarrow \Gamma \cdot (c-1)+1$
    $t_{end} \leftarrow \min(N_{test}, \Gamma \cdot c)$
▷ Compute the error for each of the $M$ algorithms
    for $i \leftarrow 1$ to $M$ do
        $\hat{\mathbf{R}}^{(c,i)}_{test} \leftarrow m^{(k,i)}(\mathbf{P}^{(k)}_{trn}, \mathbf{R}^{(k)}_{trn}, \mathbf{P}^{(k)}_{test}[t_{start} : t_{end}, \ ])$
▷ $m^(k,i)$ produces the outputs of the $i^{th}$ algorithm, trained with input $\mathbf{P}^{(k)}_{trn}$ and output $\mathbf{R}^{(k)}_{trn}$ on the testing set $\mathbf{P}^{(k)}_{test}$
        $\mathbf{E}^{(c,i)} \leftarrow \hat{\mathbf{R}}^{(c,i)}_{test} - \mathbf{R}^{(k)}_{test}[t_{start} : t_{end}, \quad ]$
        $\mathbf{E}^{(c,i)} \leftarrow \langle \mathbf{E}^{(c,i)}, \mathbf{E}^{(c,i)} \rangle$
▷ Update the weights for each of the $F = N * H$ maps
        for $j \leftarrow 1$ to $F$ do
            $\mathbf{w}^{(c,i)}[j] \leftarrow 1/mean(\mathbf{E}^{(c,i)}, j)$
▷ $mean(\mathbf{E}, j)$ computes the mean of the $j^{th}$ column of the $\mathbf{E}$ matrix
        end for
    end for

▷ Dynamic Adaptive Algorithm Combination
    for $j \leftarrow 1$ to $F$ do
        for $i \leftarrow 1$ to $M$ do
▷ Normalize weights for the $j^{th}$ variable
            $\mathbf{w}^{(c,i)}_{norm}[j] \leftarrow \mathbf{w}^{(c,i)}[j] / \sum_i^M \mathbf{w}^{(c,i)}[j]$
▷ Combine normalized weights using FF vector $\mathbf{\Lambda}$
            $\mathbf{w}^{(c,i)}_{norm}[j] \leftarrow \sum_{v=1}^{V} \mathbf{\Lambda}[v] \mathbf{w}^{(c-v,i)}_{norm}[j]$
        end for
        $\hat{\mathbf{X}}[t_{start} : t_{end}, j] \leftarrow \sum_{i=1}^{M} \mathbf{w}^{(c-1,i)}_{norm}[j] \hat{\mathbf{R}}^{(c,i)}_{test}[\quad , j]$
    end for

▷ Sliding window approach to keep last $V$ weight matrices $\mathbf{w}$ and discard the oldest one
    for $i \leftarrow 1$ to $M$ do
        for $v \leftarrow 1$ to $V$ do
            $\mathbf{w}^{(c-v,i)}_{norm} \leftarrow \mathbf{w}^{(c-v+1,i)}_{norm}$
        end for
    end for
end for
```

practice, Γ and Λ_v values are set by considering a grid search procedure over a training portion of the historical series.

IV. Experimental Assessment

A. Experimental Configuration

We consider two real wind power forecasting case studies: one based on a public dataset and the other based on a proprietary dataset. The first dataset includes 22 eastern Australian wind farm power generation time series with 5 minutes time resolution for 1 year, which is scaled to 15 minutes, the data are available with the R package of [12]. The latter is a private domain dataset that contains 28 southern Italian wind farms with a 15 minutes time resolution for 1 year. For the sake of conciseness, the tests are conducted by considering a forecasting horizon span ranging from 1 to 3 hours ($\mathcal{H} = \{1, \dots, 12\}$ time step ahead).

A standard procedure for the assessment of the forecasting model performance considers splitting the initial dataset into training and validation sets. Thus, a forecasting model uses the training set to learn, and the validation set to assess its accuracy using unknown data. Unfortunately, this approach shows several shortcomings [27]. Particularly, it neglects the impact of outliers on the models and limits the generalization capability of the model over the time.

Differently, the rolling window procedure splits the original dataset into K parts (trials). Thus, each part is further split in training and validation sets. Successively, a statistical analysis is performed on the collected accuracy metrics to analyze the behavior of the model across time. Particularly, the number of samples is 300 (10 days) and 100 (4 days) for each k-th training and validation sets, respectively. Both datasets are split in $K = 45$ trials.

The *DAFT-E* model includes Random Forest (*RF*) [28], Lazy Learning (*Lazy*) [29], and Persistence (*Naive it*). These models are selected after a preliminary analysis due to both their low computational burden and the good generalization performance, which were shown in different real-world time series forecasting problems [30]. Furthermore, the ML models are trained with different feature sets for enhancing the generalization capability of the ensemble model. The first set considers "*raw*" features, whereas the latter consider "*raw and derived*" ones.

We expect that this increases the model generalization capability, since the final forecasting is obtained by MSE-based weighted average considering the dynamic model performance over the time. Particularly, the weights are updated at the end of a whole forecasting horizon span $\Gamma = H$, where the ML are parallelly processed by mRMR [26] by extracting a feature subsets of $N_S = 5$ from a $N' \approx 2000$ feature set. The *DAFT-E* employs $V = 3$ FFs, where the weight values are $\Lambda_1 = 0.5$, $\Lambda_2 = 0.35$, and $\Lambda_3 = 0.15$. The optimal parameters are chosen after a grid search analysis on a subset of the original data.

B. Performance Validation Metric

Validation is a crucial step in the assessment of a forecasting model but in literature the focus is typically on the average prediction accuracy (e.g. Mean-Squared-Error) disregarding other relevant aspects for the decision maker. Here we propose a more general approach to assess the performance accounting for other performance measures like the spatial spread analysis and the tail error distribution analysis.

1) Spatial performance analysis: The spatial analysis of forecasting performance is important to assess possible distortions in studies that employ the obtained predictions, since an unbalanced forecasting performance may compromise the quality of the decision making process.

A reliable multivariate wind power forecasting methodology should assure a consistent prediction performance across measurement points. Inconsistent performance may compromise the quality of the decision-making process based on the obtained predictions.

Traditionally, the prediction accuracy of a multivariate forecasting methodology is assessed by computing a single error metric over all the target variables.

Unfortunately, this approach neglects the distribution of the forecasting model errors across the target variables. Therefore, a novel validation procedure is proposed to fulfill this gap inspired by the modern portfolio theory [31].

In particular, the following quantities are computed:

- $\mu_{\text{ERR}}^{(\omega,k,h)}$: the average value of the considered error metric ERR across the target variables for the forecasting model ω, forecasting horizon h and trial k.
- $\sigma_{\text{ERR}}^{(\omega,k,h)}$: the standard deviation of the considered error metric across the target variables for the forecasting model ω, forecasting horizon h and trial k;

Finally, a bivariate distribution is built by collecting $\mu_{\text{ERR}}^{(\omega,k,h)}$ and $\sigma_{\text{ERR}}^{(\omega,k,h)}$ for each trial in the experiment.

The results are visualized by means of a bivariate extension [18] of the univariate box-plot where the conventional Interquartile Range and whiskers are replaced by two convex hull polytopes. In particular, the 50% of population is included in dark-colored area (*bag*), the 99.7% in the the light-colored one (*fence*), and the points outside the fence are considered outliers (Fig. 3,4). The position of the bag and fenced area on the plot indicate the absolute performance while the size of the areas is an indicator of the uniformity of the performance across the wind farms.

2) Robustness Assessment:

a) Tail Analysis of Wind Power Forecasting Error: Although the spatial analysis of the performance returns a clear picture of the forecasting model behavior, it does not supply enough information about the worst-case configuration. To address such aspect, we adopt the tail analysis of the forecasting error distribution by using some well-known metrics in the financial domain: the value at risk (VaR), and the conditional value at risk (cVaR). Traditionally, these metrics summarize the probability distribution of financial returns to determine the worst-case financial losses given a risk threshold [32] and a strategy. In this manuscript, we apply them to the probability distribution of the forecasting errors, where each forecasting model corresponds to a different strategy. In other words, we are assessing, given an identical risk threshold for all forecasting models, which one returns the least absolute error.

In order to maintain a coherent interpretation with respect to the financial counterparts, the analysis considers absolute errors.

The rationale is twofold: first, any gap between predicted and actual value is detrimental regardless of the sign; second, we do not convert the forecasting error into TSO economic losses since this would require a deeper analysis (and more complex simulations) which are out of the scope of the manuscript.

Although it relies on some simplifying hypotheses, this approach is crucial to assess the reliability of the forecasting model. Indeed, it is reasonable to assume that the lower is the risk to commit a large prediction error, the lower is the possible associated economic losses.

Algorithm 2 Algorithm for MAE,VaR, and cVaR estimation for the ω-th model

```
for $h \in \{1,\dots,H\}$ do
    for $k \in \{1,\dots,K\}$ do
        Compute the absolute error matrix
        Compute MAE$^{(\omega,k,h)}$,VaR$^{(\omega,k,h)}$,and cVaR$^{(\omega,k,h)}$
    end for
    ▷ A $K$ MAE, VaR, and cVaR value collection is obtained, hence the statistical quantities are computed again on the obtained distributions
    $\overline{\text{MAE}}^{(\omega,h)} \leftarrow \text{mean}(\{\text{MAE}^{(\omega,1,h)},\dots,\text{MAE}^{(\omega,K,h)}\})$
    $\overline{\text{VaR}}^{(\omega,h)} \leftarrow \text{VaR}_\alpha(\{\text{VaR}^{(\omega,1,h)},\dots,\text{VaR}^{(\omega,K,h)}\})$
    $\overline{\text{cVaR}}^{(\omega,h)} \leftarrow \text{cVaR}_\alpha(\{\text{cVaR}^{(\omega,1,h)},\dots,\text{cVaR}^{(\omega,K,h)}\})$
end for
```

Algorithm 2 summarizes the procedure to estimate VaR and cVaR from the experimental results, given the ω-th forecasting model, and the h-th forecasting horizon, where

$$\text{VaR}_\alpha(X) := \min\{z|\widehat{F}_X(z) \geq \alpha\} \tag{11}$$

$$\text{cVaR}_\alpha(X) := \mathbb{E}[X|X \geq \text{VaR}_\alpha(X)] \tag{12}$$

where X is the absolute error distribution, which is obtained collecting forecasting error across all the wind farms, given the ω-th" model and h-th forecasting horizon, $\widehat{F}_X$ is the empirical cumulative distribution function, and α is the confidence level. In the results, VaR and cVaR are compared to the average value of X that is the Mean Absolute Error (MAE) (Fig. 1).

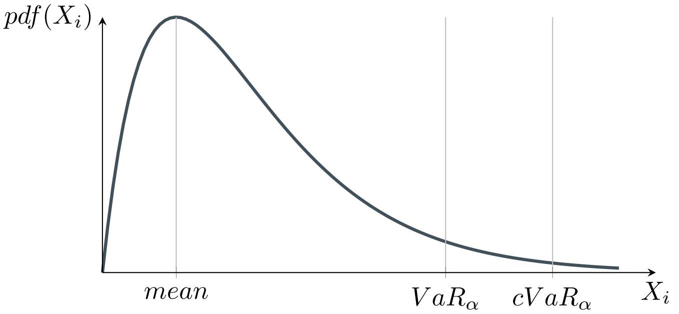

Fig. 1: Qualitative visualization of the expected value (mean), VaR_α, and $cVaR_\alpha$ for a generic non-negative loss distribution.

It should be noted that VaR is the maximum value the decision-maker accepts to lose in a percentage of the cases equal to α, as shown in (11). Mathematically, the VaR is equivalent to the α-th percentile of an empirical distribution.

The smaller is this value, the smaller is the maximum absolute forecasting error we accept to commit by using a forecasting model in the α % of cases. In other words, the lower the VaR, the more reliable the model is.

Unfortunately, VaR is not such a good metric of risk since it neglects the loss values greater than VaR. Differently, cVaR is a coherent and robust measure of risk. Hence, it supplies information about the expected losses greater than VaR as shown in (12). Particularly, cVaR is computed through the Convex Combination Formula [32]. In other terms, the cVaR represents the upper bound for the worst expected forecasting error.

b) Current Limits and Future Developments: A possible improvement of the proposed analysis could relate the forecasting errors to the impact of the decisions made by the operators on the basis of the forecasts [33]. For example, an operator may be induced to buy excess reserves or up-regulation if the forecasts exceed the reality, or, on the other way round, lines may be overloaded and generation could exceed safety margins.

A future direction of our research will be oriented toward the conceptualization of decision support systems, which aim at defining optimal bidding strategies of reserve capacity on ancillary service markets on the basis of multiple forecasted profiles.

The main idea is to apply the proposed methods in the task of forecasting the wind power production profiles, the market prices and the power network congestions in order to identify the bidding strategies that minimize the expected cost of reserve procurement. To this aim, we are currently developing stochastic optimization models, which estimate the probability distribution of the forecasting errors in order to generate a comprehensive set of scenarios. These models might also be used to estimate the economic impacts derived by the employment of the proposed forecasting models in a realistic operation scenario.

C. Benchmarks

The DAFT-E approach is benchmarked against a number of state-of-the-art techniques, which are summarized in Table I.

1) Univariate Techniques: Univariate approaches may be used to tackle a multivariate forecasting task by decomposing it in N SISO tasks or N MISO tasks.

SISO approaches typically rely on statistical approaches like Naive and Holt models [34]. The Naive is a random walk model, assuming that future values will be the same as that of the last known observation, while Holt performs an exponential smoothing [35] of the data (controlled by the parameter α_{HW}, fitted from the available data), followed by an extrapolation assuming a linear trend. Despite its trivial nature, in real-world tasks the Naive method often outperforms much more complex learning strategies: for that reason it will be considered as a baseline to normalize all our accuracy results in Section V.

MISO approaches typically use supervised learning approaches [36]. In the experimental session we considered a lazy learning (k-nearest neighbors technique [29], [37]) and an ensemble based technique (random forest [38]). In a lazy learning technique the learning process from the data is delayed until prediction time, while the k-nearest neighbor aspect allows performing a prediction of the future values given the k past points most similar to the prediction candidate. Our choice of a k-nearest neighbors lazy learning technique is motivated by two reasons: the reduced computational cost and the capability of the model to exploit local patterns in the data. The optimal value of k employed for the predictions, is automatically determined according to the input data supplied to the method. On the other hand, an ensemble technique combines the predictions coming from different base models [39], in order to improve the forecasting accuracy of the individual models and reduce the variance of the prediction. A random forest is constructed by combining several individual models (N_{RF} decision trees, whose number is optimized during the learning process), with a bagging procedure (i.e. training each model on a different subset, uniformly sampled with replacement from the original dataset). In order to assess the impact of the feature engineering and dimensionality reduction methods, these categories of models have been tested on the raw data, as well as the feature augmented, embedded data, both with and without PCA.

2) MIMO Techniques: MIMO approaches aim to capture in a single model both the temporal and cross-series (e.g. spatial) dependencies between time series. In our experiments, we considered models assuming a linear dependence between the time series (VAR, in its standard form [8] and in a state-of-the-art approach with adaptive regularization [12]) as well as a non-linear dependence among the variables (with an end-to-end recurrent neural network).

a) VAR: A VAR model describes the evolution of the multivariate time series as a linear function of their past values. A VAR model is characterized by its model order L ($L = 5$ for our experiments), denoting how many values from the past of the time series should be taken into consideration for the forecast, VAR and state space models have been shown to be equivalent and their equivalence is discussed in [40].

For a VAR model to be employed, data must meet some conventional requirements (e.g. stationarity). Also they are not suitable for high-dimensional time series data since a VAR model of N attributes with an embedding order equal to L has at least LN^2 parameters (L A_k matrices). This number of variables can be handled in the case of small problems which involve only a moderate number of attributes (i.e. N smaller than 20). In order to deal with the dimensionality issue, we included in our experiments *VARon it*, an optimized version of a VAR(L) model, where the number of parameters is reduced by means of a LASSO regularization (controlled by the parameter λ_{VAR}, optimized on the input data), and the model parameters are updated by means of an online procedure with a reduced computational cost [12].

b) Recurrent Neural Networks: Recurrent Neural Networks (RNN) is a state-of-the-art neural network family of approaches where recurrent connections between nodes allow the modeling of dynamic temporal dependencies.

For our experiments, we considered a RNN, employing one hidden layer of LSTM (*Long-Short Term Memory cells* [41]) cells, a specific type of neurons optimized for modeling temporal dependencies and for faster training. Despite these advances, the effective training of RNN remains a challenging task, due to the high number of hyperparameters to tune (i.e. layers, cells per layer, dropout, regularization). In our experiments we fixed the hidden layer number to one, and

TABLE I: Characteristics of the considered Wind Power Forecasting Models.

Model	Type	Architecture	Parameters	Feature Engineering	Embedding	Feature Selection	Cardinality Reduction	Strategy
Dynamic Adaptive Feature Temporal Ensemble – DAFTE	Hybrid Ensemble	N MISO	V, Λ	(Yes)	(Yes)	(Yes) - MRMR	(No)	Direct
Naive it – Persistence (*)	Naive	N SISO	$\emptyset$	(No)	(No)	(No)	(No)	Iterative
Random Forest FS all (*)	Machine Learning	N MISO	N_{RF}	(Yes)	(Yes)	(Yes) - MRMR	(No)	Direct
Random Forest FS raw (*)	Machine Learning	N MISO	N_{RF}	(No)	(Yes)	(Yes) - MRMR	(No)	Direct
Lazy Learning FS all (*)	Machine Learning	N MISO	k	(Yes)	(Yes)	(Yes) - MRMR	(No)	Direct
Lazy Learning FS raw (*)	Machine Learning	N MISO	k	(No)	(Yes)	(Yes) - MRMR	(No)	Direct
Holt-Winter Exponential Smoothing – Es it	Statistical	N SISO	α_{HW}	(No)	(No)	(No)	(No)	Iterative
Lazy PCA	Machine Learning	N MISO	k	(Yes)	(Yes)	(No)	(Yes) - PCA	Direct
Long Short Term Memory RNN - LSTM	Deep Learning	MIMO	N_{LSTM}, δ_{LSTM}	(No)	(No)	(No)	(Yes) - PCA	Direct
Random Forest PCA	Machine Learning	N MISO	N_{RF}	(Yes)	(Yes)	(No)	(Yes) - PCA	Direct
Var it	Statistical	MIMO	L	(No)	(No)	(No)	(No)	Direct
Online Var	Statistical	MIMO	L, λ_{VAR}	(No)	(No)	(No)	(No)	Direct

(*) *internal algorithms of* ***DAFT-E***

we performed a grid search, on the test set, over different values of cells per layer, dropout rate in the input and recurrent elements, and regularization technique (no regularization, L_1, L_2, and a combination of both). The resulting architecture employs $N_{LSTM} = 100$ cells, no regularization and a dropout rate $\delta_{LSTM} = 0.2$ for both the input and recurrent elements. It should be noted that using the test set for the model selection might yield over-optimistic performance. Similar architectures have been considered as state-of-the art techniques in a recent survey [14].

V. Experimental Results

The experiments were run on a shared infrastructure equipped with Intel Xeon E5-2640 V4 – 10 core CPU and Asus GTX 1080 TI GPU. All the forecasting models have been implemented with the same programming language (R 3.6), employed in serial mode (on a single core /single GPU, in case of deep learning models), with the same limitation in terms of RAM (5GB of maximum available memory). Figs. from 2 to 6 visualize the experimental results of the two benchmarks. Since the Australian benchmark is public (unlike the Italian), due to space restrictions, we decided to allocate more space to the figures that illustrate the Australian results.

The MSE reduction ratio (nMSE) between the ω-th model and the Naive baseline for the k-th trial and the h-th forecasting horizon is computed as:

$$\text{nMSE}^{(\omega,k,h)} = (\text{MSE}^{(\omega,k,h)} / \text{MSE}^{(Naive,k,h)}) - 1 \qquad (13)$$

Fig. 2 shows the distribution of the nMSE according to (13) across K trials and for different forecasting horizons. Note that only negative values of nMSE correspond to an improvement in accuracy with respect to the Naive baseline.

Figs. 3 and 4 show the bag-plots of the bivariate distribution μ_{MSE}-σ_{MSE} obtained for the Australian and Italian benchmark, respectively. The closer is the cloud point to the lower-left corner (low mean - low variance of the forecasting error), the better is the performance of the ω-th forecasting model over the K-th trials. Note that in this visualisation the variability over the vertical (horizontal) axis is related to the variability across (within) trials. Fig. 3 shows that the DAFT-E combination strategy outperforms the single components (Lazy FS all, RF FS all, RF FS raw) taken individually, as well as state-of-the-art approaches. We get a similar result for the Italian benchmark (Fig. 4) but, for the sake of space, we report only a smaller set of approaches.

Fig. 5 shows the risk measures (MAE, VaR (11) and cVaR (12)) across different forecasting horizons. This representation allows to compare the average performance of different forecasting techniques (MAE) vs the worst-case ($\alpha = 95\%$) configurations. In particular, we consider that the lower the increment of those quantities for increasing horizons, the higher is the robustness.

Finally, Fig. 6 shows the computational times of all models. The total computation time accounts for feature engineering, embedding, feature selection, and model training steps. If a model does not include some of these steps, the corresponding computational time is null.

Overall, some general considerations may be made on the basis of the experimental results:

1) Multi-step-ahead wide area wind power forecasting relying solely on historical power data is a challenging task, as shown by the difficulty in improving over simpler baselines (Naive, Holt-Winters).
2) The proposed approach *DAFT-E* is a promising alternative to the state-of-the-art forecasting strategies in this context. In terms of nMSE, DAFT-E is the best model, followed by *VARon it* (Fig. 2). The *LSTM dir* accuracy is definitely the worst one. The *ES* performance dramatically decreases as h increases;
3) The *DAFT-E* has a balanced performance on the plane μ_{MSE}-σ_{MSE} over h as shown by Fig. 3. This is shown by the fact that its bagplot area is the smallest and is closer to the origin with respect to those of multivariate approaches like *VARon it*;
4) The *DAFT-E* has the least absolute error in terms of VaR and CVaR, for a $\alpha = 95\%$ confidence level. In particular, the best *DAFT-E* is the one using *RF FS all* and *RF FS raw* (Fig. 5);
5) The addition of derived features to the information set improves the accuracy compared to the adoption of raw features only, particularly for large h (Fig. 3);
6) *DAFT-E* demands a much larger computational time than the fastest method (*VARon it*), yet such overhead is compensated by a better performance accuracy (Fig. 6). The same reasoning does not apply to *LSTM dir* whose heavier computational time is not compensated by an increase in accuracy;
7) The embedding procedure covers about 50% of the entire computational time. Future work will focus on speeding up this step (e.g. by making it parallel).

VI. Conclusions

Erratic wind behavior may engender critical issues during the grid operation. To ensure safety it is then essential to develop effective and reliable models relying on the minimal

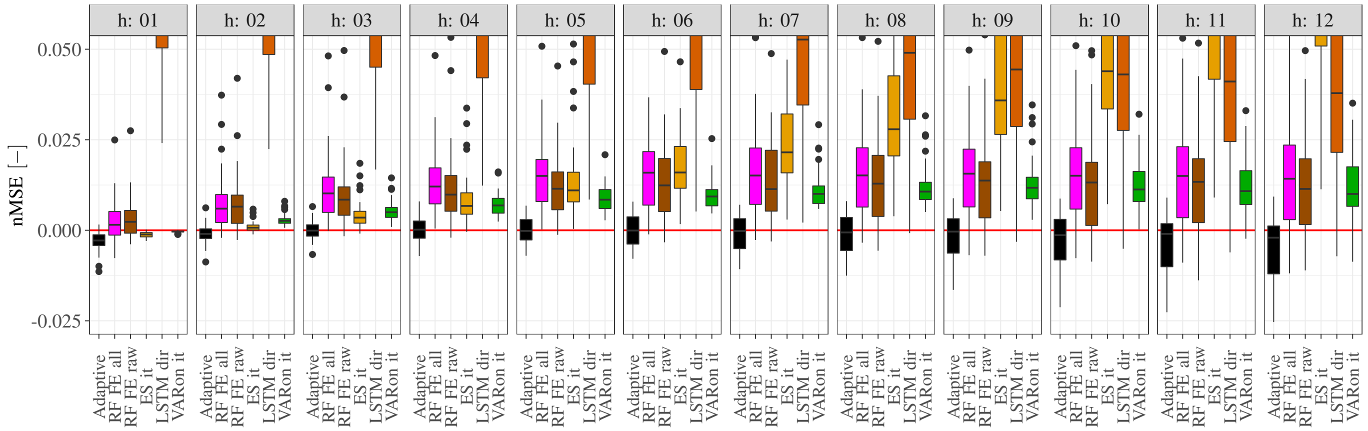

Fig. 2: Visualization of the nMSE across the forecasting horizon for the Australian case study. The lower the nMSE is, the better the corresponding model is performing. An nMSE < 0 indicates that the corresponding model is outperforming the Naive model. The plot includes the Proposed Model (DAFT-E), the best performing algorithms among those combined in DAFT-E (RF FS all, RF FS raw), and the benchmark models showing the best performance. The same figure with greater vertical range scale is included into the supplementary material. The forecasting horizon for each panel is equal to $h \cdot$ 15 min.

amount of data made available from the remote terminal units of the grid (wind power injections).

This paper proposes a machine-learning based model for wind power forecasting (DAFT-E), which deploys dynamic adaptive ensemble and feature extraction techniques. It achieves a lower overall MSE or MAE than other state-of the art time series methods, while also lowering variability of performance among sites (geographically) and over time (different forecast periods) and also lowering the frequency (risk) of large errors Moreover, the proposed multivariate-to-univariate problem decomposition approach is well-suited for parallel computation. In combination with an efficient feature selection process, this approach could easily scale up to larger multivariate problems. This study supports the idea that, despite the success of deep learning representation learning strategies in a number of applied settings (e.g. image classification), we should not expect any “a priori” superiority of such methods, notably in power systems settings where expert knowledge is relevant and accurate.

Further studies will focus on further exploration of automatic feature selection and algorithm choice for the ensemble, as well as the usage of multivariate models in order to exploit the spatio-temporal correlation among the wind farms.

Acknowledgment

We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Gianluca Bontempi was supported by Service Public de Wallonie Recherche under grant n° 2010235–ARIAC by DIGITALWALLONIA4.AI.

References

[1] A. Soroudi and T. Amraee, “Decision making under uncertainty in energy systems: State of the art,” *Renewable and Sustainable Energy Reviews*, vol. 28, pp. 376–384, 2013.

[2] C. Feng, M. Sun, M. Cui, E. K. Chartan, B.-M. Hodge, and J. Zhang, “Characterizing forecastability of wind sites in the united states,” *Renewable Energy*, vol. 133, pp. 1352–1365, 2019.

[3] B.-M. Hodge and M. Milligan, “Wind power forecasting error distributions over multiple timescales,” in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–8.

[4] M. Zou, D. Fang, S. Djokic, V. Di Giorgio, R. Langella, and A. Testa, “Evaluation of wind turbine power outputs with and without uncertainties in input wind speed and wind direction data,” *IET Renewable Power Generation*, vol. 14, no. 15, pp. 2801–2809, 2020.

[5] S. Hanifi, X. Liu, Z. Lin, and S. Lotfian, “A critical review of wind power forecasting methods—past, present and future,” *Energies*, vol. 13, no. 15, p. 3764, 2020.

[6] C. Ning and F. You, “Data-driven adaptive robust unit commitment under wind power uncertainty: A bayesian nonparametric approach,” *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 2409–2418, 2019.

[7] J. S. Roungkvist and P. Enevoldsen, “Timescale classification in wind forecasting: A review of the state-of-the-art,” *Journal of Forecasting*, 2020.

[8] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[9] P. Exterkate, P. J. Groenen, C. Heij, and D. van Dijk, “Nonlinear forecasting with many predictors using kernel ridge regression,” *International Journal of Forecasting*, vol. 32, no. 3, pp. 736–753, 2016.

[10] J. De Stefani, Y.-A. Le Borgne, O. Caelen, D. Hattab, and G. Bontempi, “Batch and incremental dynamic factor machine learning for multivariate and multi-step-ahead forecasting,” *International Journal of Data Science and Analytics*, vol. 7, no. 4, pp. 311–329, 2019.

[11] L. Cavalcante, R. J. Bessa, M. Reis, and J. Browell, “Lasso vector autoregression structures for very short-term wind power forecasting,” *Wind Energy*, vol. 20, no. 4, pp. 657–675, 2017.

[12] J. W. Messner and P. Pinson, “Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting,” *International Journal of Forecasting*, vol. 35, no. 4, pp. 1485–1498, 2019.

[13] Y. Zhao, L. Ye, P. Pinson, Y. Tang, and P. Lu, “Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting,” *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5029–5040, 2018.

[14] X. Deng, H. Shao, C. Hu, D. Jiang, and Y. Jiang, “Wind power forecasting methods based on deep learning: A survey,” *Computer Modeling in Engineering & Sciences*, vol. 122, no. 1, pp. 273–302, 2020.

[15] M. Khodayar and J. Wang, “Spatio-temporal graph deep neural network for short-term wind speed forecasting,” *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 670–681, 2019.

[16] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, “A review of deep learning models for time series prediction,” *IEEE Sensors Journal*, 2019.

[17] T. Guo, Z. Xu, X. Yao, H. Chen, K. Aberer, and K. Funaya, “Robust online time series prediction with recurrent neural networks,” in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Ieee, 2016, pp. 816–825.

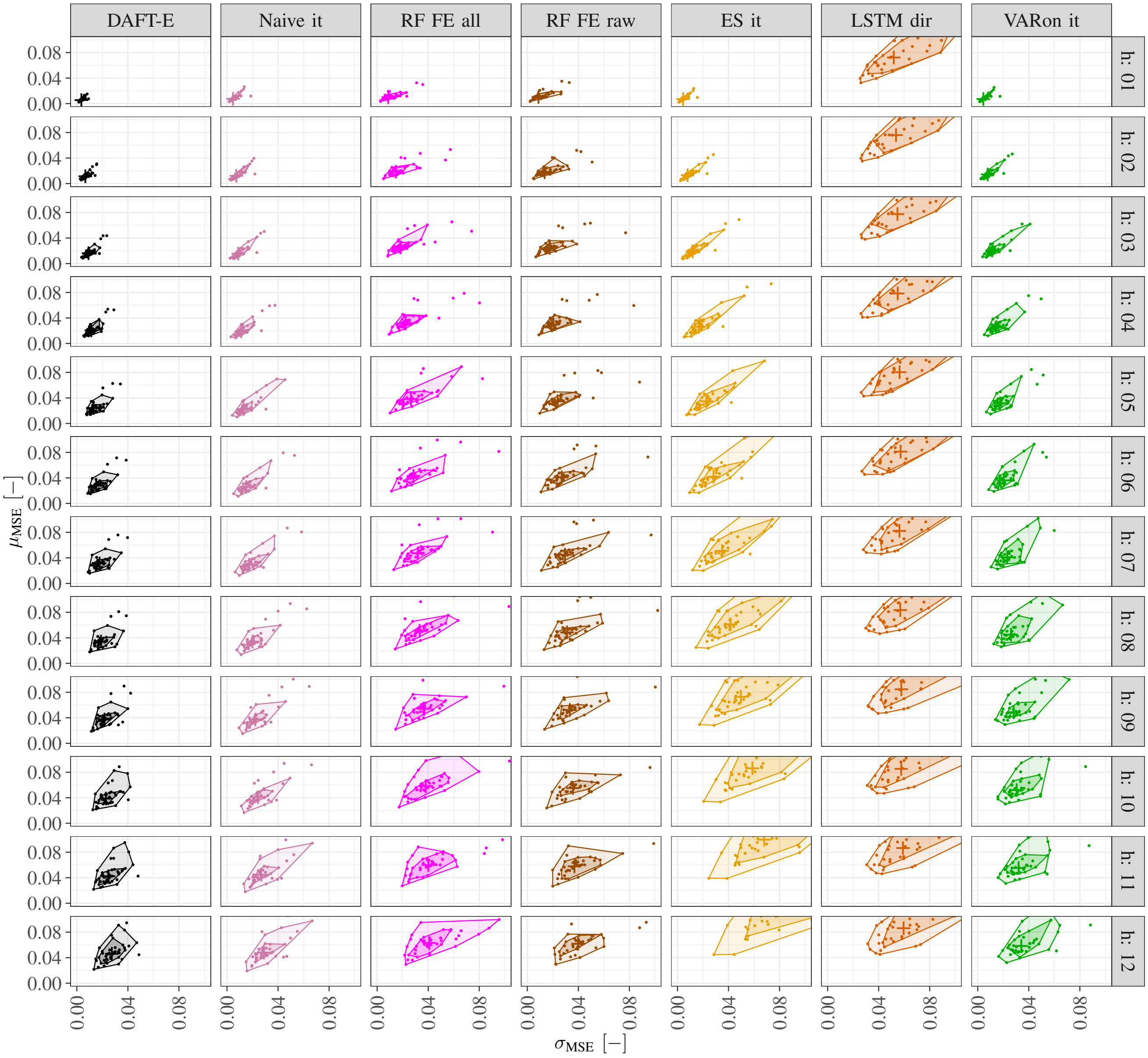

Fig. 3: Visualization of the bagplot of the bivariate distribution μ_{MSE}-σ_{MSE} across the forecasting horizon h for the proposed model (DAFT-E), the internal algorithms of DAFT-E (Lazy FS all, RF FS all, RF FS raw, Naive it), and the benchmark models showing the best performance - Australian case study. A smaller bagplot area indicates a reduced variability in the corresponding method's predictions. The forecasting horizon for each panels row is equal to $h \cdot$ 15 min.

[18] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: a bivariate boxplot," *The American Statistician*, vol. 53, no. 4, pp. 382–387, 1999.

[19] T. Januschowski, J. Gasthaus, Y. Wang, D. Salinas, V. Flunkert, M. Bohlke-Schneider, and L. Callot, "Criteria for classifying forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 167–177, 2020.

[20] H. Liu and C. Chen, "Data processing strategies in wind energy forecasting models and applications: A comprehensive review," *Applied Energy*, vol. 249, pp. 392–408, 2019.

[21] S. B. Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa, "A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition," *Expert systems with applications*, vol. 39, no. 8, pp. 7067–7083, 2012.

[22] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[23] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro *et al.*, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific reports*, vol. 5, no. 1, pp. 1–12, 2015.

[24] F. De Caro, J. De Stefani, G. Bontempi, A. Vaccaro, and D. Villacci, "Robust assessment of short-term wind power forecasting models on multiple time horizons," *Technology and Economics of Smart Grids and Sustainable Energy*, vol. 5, no. 1, pp. 1–15, 2020.

[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[26] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "mrmre: an r package for parallelized mrmr ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, 2013.

[27] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International journal of forecasting*, vol. 16, no. 4, pp. 437–450, 2000.

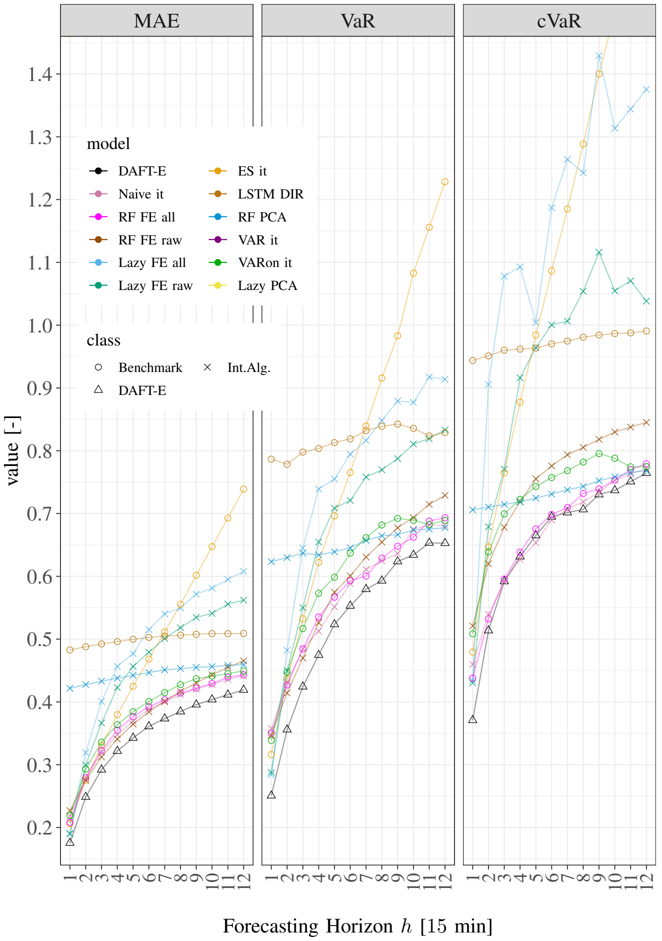

Fig. 5: Visualization of the expected value (mean), VaR_α, and $cVaR_\alpha$ ($\alpha = 95\%$) across the forecasting horizon for all models - Australian case study. For all the three metrics, the lower the metric value is, the smaller is the risk (in terms of absolute forecasting error) that we are going to expect using the corresponding method.

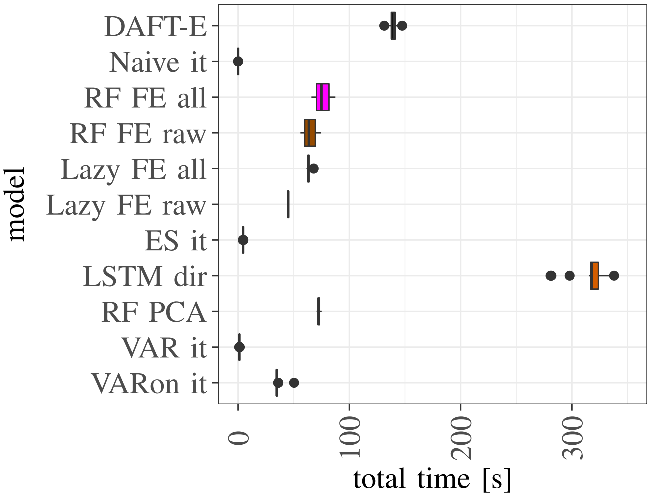

Fig. 6: Spread of the computational time across K trials, Australian Case Study. Smaller computational times indicate an higher computational efficiency of the methods.

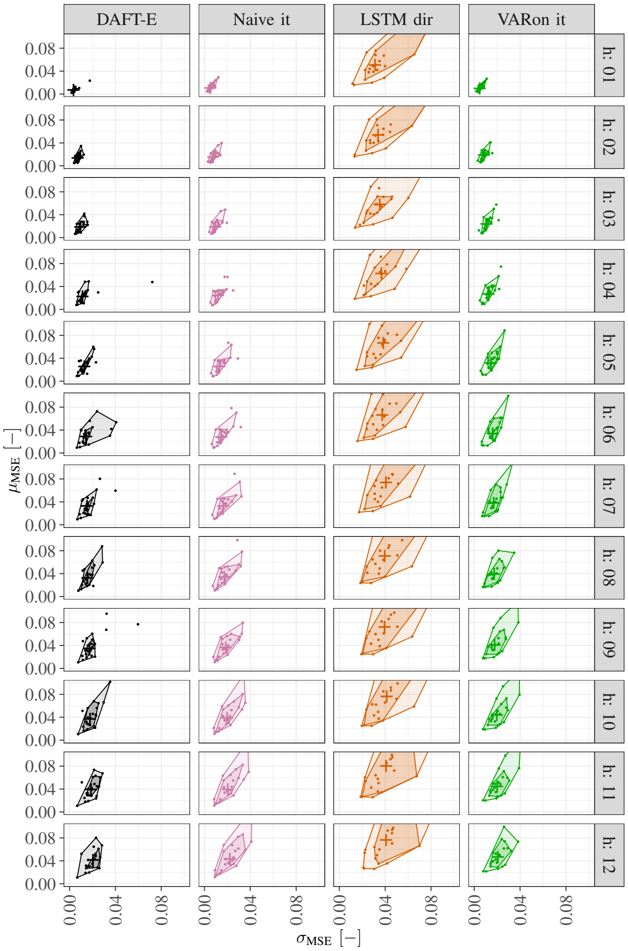

Fig. 4: Visualization of the bagplot of the bivariate distribution μ_{MSE}-σ_{MSE} across the forecasting horizon h for the proposed model (DAFT-E), and the benchmark models showing the best performance - Italian case study. A smaller bagplot area indicates a reduced variability in the corresponding method's predictions. The forecasting horizon for each panels row is equal to $h \cdot$ 15 min.

[28] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[29] G. Bontempi, M. Birattari, and H. Bersini, "Lazy learning for local modelling and control design," *International Journal of Control*, vol. 72, no. 7-8, pp. 643–658, 1999.

[30] A. R. S. Parmezan, V. M. Souza, and G. E. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Information sciences*, vol. 484, pp. 302–337, 2019.

[31] H. Markowitz, "Portfolio selection*," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1952.tb01525.x

[32] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[33] G. Li and H.-D. Chiang, "Toward cost-oriented forecasting of wind power generation," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2508–2517, 2016.

[34] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The m4 competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020.

[35] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International journal of forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[36] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in *European business intelligence summer school*. Springer, Berlin, Heidelberg, 2012, pp. 62–77.

[37] G. Bontempi and S. B. Taieb, "Conditionally dependent strategies for multiple-step-ahead prediction in local learning," *International journal of forecasting*, vol. 27, no. 3, pp. 689–699, 2011.

[38] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] J. M. Bates and C. W. Granger, "The combination of forecasts," *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451–468, 1969.

[40] P. Gilbert., "State space and ARMA models: an overview of the equivalence," 1993.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Fabrizio De Caro (S'17, M'21) received the B.Sc and M.Sc in Energy Engineering (2014 and 2016), and Ph.D. degree in Information Technologies for Engineering (2020) from the University of Sannio, (UniSannio), Benevento, Italy. He is currently a Post-Doc Researcher with the Power System Research Group (PSRG) at the Engineering Department, UniSannio. His research interests include the effective renewable energy sources integration in smart grids, wind power forecasting, artificial intelligence in power systems, decision-making tools in the presence of uncertainty, electricity markets, and resilience of power systems. He is Associate Editor of the *Technology and Economics of Smart Grids and Sustainable Energies*.

Jacopo De Stefani received the M.Sc. degrees in Computer Science and Engineering from the Université Libre de Bruxelles, Bruxelles, Belgium, in 2013 and in Engineering of Computing Systems from the Politecnico di Milano, Milano, Italy, in 2015. He is currently a PhD candidate at Université Libre de Bruxelles, Bruxelles, Belgium. Since October 2021, he is a Lecturer at the Technology, Policy and Management faculty at Delft University of Technology (TUDelft). His research interests include multivariate time series analysis, multiple-step-ahead forecasting, renewable energy forecasting and sustainable computing.

Alfredo Vaccaro Alfredo Vaccaro (Senior Member, IEEE) received the M.Sc. (Hons.) degree in electronic engineering from the University of Salerno, Salerno, Italy, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada. From 1999 to 2002, he was an Assistant Researcher with the Department of Electrical and Electronic Engineering, University of Salerno. From March 2002 to October 2015, he was an Assistant Professor of electric power systems with the Department of Engineering, University of Sannio, Benevento, Italy, where he is currently an Associate Professor of electrical power system. His research interests include soft computing and interval-based method applied to power system analysis, and advanced control architectures for diagnostic and protection of distribution networks. Prof. Vaccaro is Associate Editor of the *IEEE trans. on Power Systems*, *IEEE trans. on Smart Grids*, and he is the Chair of the *IEEE PES Awards and Recognition Committee*.

Gianluca Bontempi is Full Professor in the Computer Science Department at the Université Libre de Bruxelles (ULB), Brussels, Belgium, co-head of the ULB Machine Learning Group (mlg.ulb.ac.be). He has been Director of (IB)2, the ULB/VUB Interuniversity Institute of Bioinformatics in Brussels (ibsquare.be) in 2013-17. His main research interests are big data mining, machine learning, bioinformatics, causal inference, predictive modeling and their application to complex tasks in engineering (time series forecasting, fraud detection) and life science (network inference, gene signature extraction). He was Marie Curie fellow researcher, he was awarded in two international data analysis competitions and he took part to many research projects in collaboration with universities and private companies all over Europe. He is author of more than 250 scientific publications and his H-number is 61. He is the Belgian (French Community) national contact point of the *CLAIRE* network, co-leader of the *CLAIRE COVID19 Task Force* and Associate Editor of the *International Journal of Forecasting*. He is also co-author of several open-source software packages for bioinformatics, data mining and prediction.