Regional Wind Power Forecasting Based On Bayesian Feature Selection

Regional Wind Power Forecasting Based On Bayesian Feature Selection

Theodoros Konstantinou, *Member, IEEE* and Nikos Hatziargyriou, *Life Fellow, IEEE*

***Abstract*—In recent years, the integration of renewable energy sources in power systems has been increasing. Their inherent unpredictability and output fluctuations pose challenges to secure power system operations and energy market pricing stability. Therefore, an accurate forecast of renewable energy generation is crucial. Several effective forecasting methods that have been applied are based on Machine Learning (ML). A key factor in the application of ML methods is the choice of input features, a task that has become more complex in regional wind power forecasting, where regions can cover entire countries. The proposed method aims to improve forecasting performance by streamlining input features through a data-driven model-agnostic preprocessing technique. This involves splitting the multidimensional numerical weather predictions into subareas and eliminating non-informative subareas. The selection of optimal split and remove parameters is guided by a Bayesian sequential optimisation process, which builds on prior knowledge from previous iterations. The approach has been tested on actual wind power measurements aggregated at the regional level for three countries located in south-east Europe and was found to be effective in improving the performance of popular data-driven forecasting methods.**

***Index Terms*—Artificial neural networks, Bayesian feature selection, numerical weather predictions, wind power forecasting**

NOMENCLATURE

BFS	Bayesian Feature Selection
D	Dimensionality of the space in which observations are defined
D_{inn}, D_{out}	Inner and Outer multidimensional spaces of possible solutions
δ	Convergence criterion threshold
EI	Expected Improvement Acquisition Function
e	Value of the Loss-Objective Function
e^*	Threshold value based on quantile γ to divide observed pairs $[e, h]$ into two subsets
$g(h)$	Density of "non-valuable" observations
$g_{inn}(h_{inn}\|h_{out})$	Density function for "non-valuable" observations for inner hyperparameters conditional to outer hyperparameters
$g_{out}(h_{out})$	Density function for "non-valuable" observations for outer hyperparameters
h	Set of hyperparameters
h^*	Candidate set of hyperparameters with the greatest expected improvement
h_{inn}	Inner vector of hyperparameters
h_{out}	Outer vector of hyperparameters
h_i	i_{th} Observation in the set of observations
IC	Installed Capacity in region under study
k	Window Function
$l(h)$	Density of "valuable" observations
$l_{inn}(h_{inn}\|h_{out})$	Density function for "valuable" observations for inner hyperparameters conditional to outer hyperparameters
$l_{out}(h_{out})$	Density function for "valuable" observations for outer hyperparameters
NMAE	Normalised Mean Absolute Error
N	Total number of observations
N_{obs}	Number of evaluation observations
N_{out}, N_{inn}	Number of samples drawn from D_{out} and D_{inn} spaces respectively
NMSE	Normalised Mean Squared Error
NWP	Numerical Weather Prediction
$n_{c,i}$	Splitting Index for Column Dimension, i denotes the index of the sub-area
$n_{r,i}$	Splitting Index for Row Dimension, i denotes the index of the sub-area
p_i	"Remove" parameters
Pr	Probability
rank	Rank of a matrix or dimension
r	Length of the window function in Parzen-window density estimation
S-R Method	Split-Remove Method
S_1, S_2	Number of "split" indices in each dimension
TPE	Tree-Structured Parzen Estimator
γ	Quantile parameter determining the threshold value e^* for dividing observations
$\hat{e_{rec}}$	Average of the errors of the five previous iterations
$\hat{y_i}$	Prediction of the i_{th} observation
y_i	i_{th} observed value

T. Konstantinou is with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece, e-mail: tkonstantinou@mail.ntua.gr.

N. Hatziargyriou is with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece and University of Vaasa, Finland, e-mail: nh@power.ece.ntua.gr

I. INTRODUCTION

WIND power plays a critical role in the transition towards a future of sustainable and clean energy. Furthermore, wind power is widely available as a domestic source

of energy; thus, it contributes to reducing energy dependence and providing a more secure energy supply. Furthermore, wind power is a cost-competitive source of energy, and thanks to its very low operating costs, it can help to stabilise energy prices, providing benefits to both energy consumers and the wider economy. As such, wind power generation is a crucial component of modern power systems.

As the share of wind power in the energy mix increases, it is increasingly important to ensure its efficient integration into power systems. An important aspect is wind power prediction with a certain level of accuracy. This information is essential for electricity market participants and system operators. Regional wind power forecasting can play a key role in the integration of wind power into large power systems by improving our understanding of the behaviour of wind energy resources and the reliability of wind power generation. Wind power forecasting, however, is a complex task that depends on various inputs and parameters, such as weather patterns, wind speed, and wind direction. Its accuracy also depends on the choice of forecasting methods and the quality of the input data. Forecasting wind power generation at the regional or country level is more complex than at the local scale, as it involves a larger number of inputs and covers a larger area of complex terrain.

A. Literature Review

The literature on wind power forecasting methods is extensive and covers a wide range of approaches, from traditional time series methods to advanced machine learning techniques. Time series methods, such as ARIMA and exponential smoothing, have been widely used for wind power forecasting due to their simplicity and ease of implementation. However, these methods are limited by their assumption of stationarity and may not be suitable for capturing complex and non-linear patterns in wind energy generation.

More recently, machine learning techniques have gained popularity in wind power forecasting due to their ability to handle complex and non-linear patterns [1]–[4]. These techniques include artificial neural networks, decision trees, support vector machines, and various deep learning models and have been shown to outperform traditional time series methods, particularly for short-term wind power forecasting. In addition, the recent advancement of big data analytics and cloud computing has enabled the application of more complex machine learning models, such as convolutional neural networks and recurrent neural networks. According to the literature, there are two principal approaches for regional wind power forecasting: the Aggregation Method and the Upscaling Method [5].

1) Aggregation Method: The Aggregation Method is employed to combine the wind power generation forecasts of individual wind farms into a single forecast for an entire region or country [6]–[9]. This is typically achieved by summing the expected individual wind farm power forecasts, taking into account the installed capacity of each wind farm. The Aggregation Method is significant for assessing wind power as it provides an overall picture of the expected wind power generation for the region. Methods to aggregate wind power forecasts include:

- Simple summation: This is the simplest method to aggregate wind power forecasts, where the forecasts for individual wind farms are simply summed to obtain the regional forecast.
- Weighted summation: In this method, a weight is assigned to each wind farm forecast based on its expected generation, and the forecasts are aggregated using the weighted summation. This method can offer improved accuracy over simple summation by taking into account the differences in wind power generation forecast errors between wind farms [10].
- Spatial interpolation:
  - Inverse Distance Weighting: Inverse Distance Weighting employs the inverse distance between the target location and the sample locations to estimate wind power values, providing a simple and computationally efficient means of spatial interpolation [11].
  - Kriging: A more sophisticated geostatistical interpolation method, Kriging utilises the spatial autocorrelation of the data to estimate values at locations where wind farms are installed but lack historical data. Kriging can capture spatial patterns of wind power generation in a large area, which helps to aggregate regional forecasts [12].

2) Upscaling Method: Upscaling methods are widely utilised in regional wind power forecasting to transfer information from high-resolution data to lower-resolution data, aiming to capture the spatial patterns and relationships of the wind power generation process in a specific region [13]–[19]. This leads to the generation of more accurate predictions at a regional level. Popular upscaling methods include:

- Machine Learning (ML): ML is highly favoured for forecasting regional and national wind power. By capturing complex non-linear relationships between the wind power generation process and its inputs, like wind speed and direction, ML models, and especially more advanced techniques, such as Deep Learning, exhibit high accuracy, especially when managing large datasets.

Unlike Aggregation methods, which aggregate data at the wind farm level to estimate the generation of wind power throughout the region, the upscaling methods are more adept at providing accurate forecasts. They use NWP and other meteorological data to model the relationship between weather conditions and wind power generation. By combining meteorological data with wind farm level observations, upscaling methods can produce regional forecasts that are both more accurate and robust to changes in wind farm configurations. Upscaling methods use higher resolution NWP data and can flexibly model complex terrain and weather conditions through advanced machine learning techniques, providing a nuanced understanding of wind power generation dynamics across regions.

Moreover, upscaling methods offer several advantages over aggregation methods. First, they can use higher-resolution NWP data and consider complex terrains, such as mountains, valleys, and coastlines, which can have a significant impact on wind speed patterns and power generation. Second, upscaling methods can model the relationship between wind power generation and weather conditions in a more flexible way, as they can incorporate advanced machine learning techniques, such as neural networks and support vector machines. These techniques can capture non-linear relationships between the inputs and outputs. Third, upscaling methods can better incorporate the effects of wind farm interactions and power grid constraints, allowing for a more accurate representation of the overall generation of wind power.

In conclusion, upscaling methods provide several advantages over aggregation methods for regional wind power forecasting, including improved accuracy, more flexible modelling, and better representation of complex relationships between weather conditions and wind power generation and are selected for the development of the proposed method.

B. Contributions

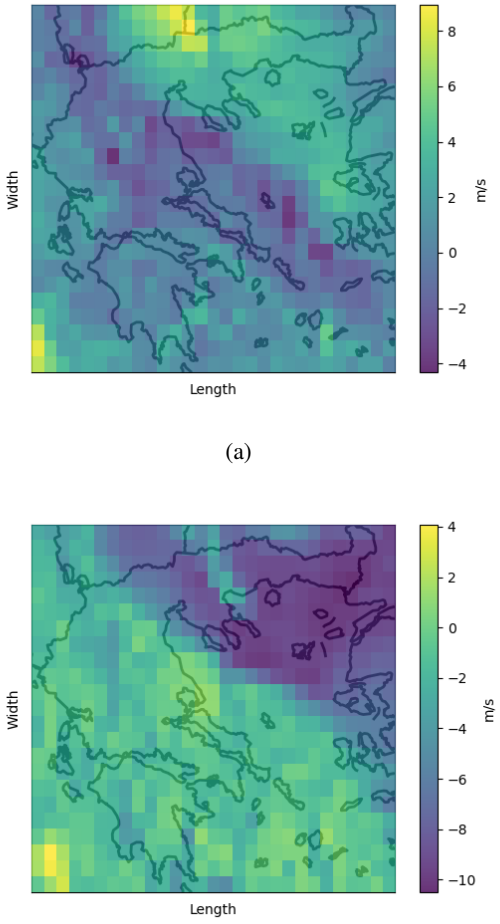
One of the challenges in using NWP for wind power forecasting, over large areas, is that these data are high-dimensional, including information from multiple variables, which may not be all relevant for the specific application. More specifically, for country-wide wind power forecasting, some sub-areas of NWP information, e.g. areas with no wind park installations, may not be directly relevant for predicting wind power generation. This can result in large amounts of irrelevant information being included in the inputs to the forecasting models, which can lead to lower accuracy and longer processing times. To address this challenge, techniques such as dimensionality reduction can be used to reduce the number of input features and eliminate less relevant information.

In this work, a dimensionality reduction method is proposed, where NWP are split into sub-areas, and only the information from the most relevant sub-areas are used as inputs to the forecasting models. The method is applied to the input data, so it is model-agnostic, i.e. it can be applied to improve the performance of any existing wind power forecasting models.

In this paper, the proposed method is tested on well-established, data-driven models such as Support Vector Machines (SVM), fully-connected artificial neural networks (ANN) and Convolutional Neural Networks (CNN) which have proven to be effective in wind power forecasting. The contribution of this paper lies in the development of a method that processes NWP as image-like data, allowing splitting of the high-dimensional data into smaller sub-areas, defined by the geographical latitude and longitude widths. After splitting, the non-informative sub-areas are removed by Bayesian feature selection. In this way, the number of input features is reduced, simplifying the forecasting model and enhancing its accuracy.

II. Numerical Weather Predictions

NWP includes multidimensional information about a variety of atmospheric and meteorological variables, such as temperature, pressure, wind speed and direction, etc. In the context of wind power forecasting, over large areas, some sub-areas of NWP data may contain information that is not representative of the wind patterns or meteorological conditions that affect the wind power generation across the region. This may be due to the influence of specific geographical or topographical features in these sub-areas, such as mountains, lakes, or other bodies of water, which create unique specific location-dependent weather conditions.



(a)

(b)

Fig. 1: (a) Spatially distributed values of the U component of the wind (10 meters above ground) (b) Spatially distributed values of the V component of the wind (10 meters above ground)

In Figure 1, the NWP values of the eastward and northward wind vectors U and V over Greece are isolated and displayed as individual images. Such image-like data are typically described by three dimensions: width, length and depth. The first two dimensions are defined by the size of the data to be modelled, while in the case of numerical weather predictions, used as image-like data, the width and length are determined by the study area and the spatial resolution of the numerical weather predictions used. The different colours in the image represent the values of each weather variable (U,V) in each isolated figure. In practise, the two figures are stacked to create

a three-dimensional matrix, where the third dimension is equal to the weather variables that are examined. As an example, the focus area for Greece is chosen between the coordinates (latitude = 36.25^o, longitude = 20.00^o) and (latitude = 41.75^o, longitude = 27.00^o), which correspond to an area equal to $(5.50^o \times 7.00^o)$ in degrees. If NWP are provided by a global weather model, which usually has a spatial resolution of 0.25 degrees, the image-like data correspond to a 23×29 matrix per weather variable at a specific altitude.

III. Bayesian Feature Selection For NWP Pre-Processing

In order to eliminate sub-areas of NWP data that are not informative or relevant for wind power forecasting over large areas, dimensionality reduction techniques are used. Such methods help to reduce the complexity of the forecasting model, which in turn leads to faster training and prediction times, increased accuracy and improved performance.

A. Feature Selection Methods

The process of eliminating spatial features from NWP can be challenging and time-consuming, in order to achieve an optimal selection of spatial features, without compromising the final results. There are several feature selection methods for input images or multidimensional data such as NWP [20]–[24]. Some of the most widely used methods are as follows:

- Filter methods evaluate the feature importance based on a set criterion and rank the features based on their scores. The most common criterion used is the correlation coefficient, which measures the linear relationship between the features and the target variable.
- Wrapper methods use the performance of a specific machine learning model as an evaluation criterion. The idea behind these methods is to train the machine learning model on a subset of features and evaluate its performance, then iteratively add or remove features to find the optimal subset.
- Embedded methods incorporate the feature selection process into the model training process itself. For example, regularisation techniques such as Lasso or Ridge regression can be used to penalise the importance of features and select only the most relevant ones.
- Bayesian Feature Selection (BFS), is a statistical approach to feature selection that utilises Bayesian learning algorithms to identify the most relevant features for a given task. BFS is based on the idea that the relevance of a feature can be estimated based on its marginal distribution in the observed data and the prior distribution of its relevance [25]–[27].

The BFS approach has several advantages over other feature selection methods, such as filter methods and wrapper methods. Unlike filter methods, BFS considers the relationships between features and the impact of removing a feature on the performance of the model. Unlike wrapper methods, BFS is computationally efficient and does not require the computation of multiple models, making it suitable for large-scale feature selection tasks. Furthermore, BFS can incorporate prior knowledge, such as domain knowledge or information from previous experiments, making it a powerful tool for feature selection in complex real-world applications. Finally, BFS is flexible and adaptable to changing circumstances, making it well suited for dynamic and complex forecasting scenarios, such as those encountered in regional wind power forecasting.

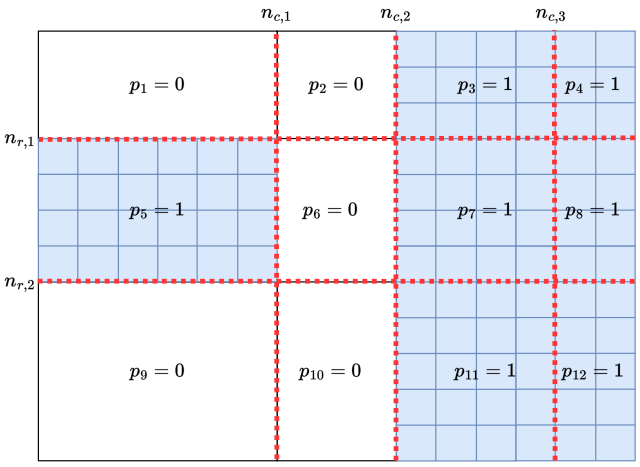


Fig. 2: Splitting and removing steps

To the authors' knowledge, a systematic elimination of sub-areas of NWP has not been studied before. In the remainder of the paper, the proposed sub-areas-based feature selection method is called the "split" and "remove" (S-R) method.

B. S-R method

The S-R method consists of two steps: First, the multidimensional data of NWP are divided in sub-areas based on the two first dimensions, which represent the 2D spatial plane, where weather predictions are given. Second, the non-informative sub-areas are discarded. For the first step, splitting indices are used to define the sub-areas, given as a vector for each of the two dimensions. These vectors, which are monotonically increasing, determine the last row or column of each sub-area in the input data matrix. In a 2×3 example, the two splitting vectors are $([n_{r,1}, n_{r,2}], [n_{c,1}, n_{c,2}, n_{c,3}])$, resulting in a split input data matrix of $(2 + 1) \times (3 + 1) = 12$ sub-areas, as shown in Figure 2. After the splitting process, a vector of binary elements, referred to as remove parameters, is generated, each corresponding to a specific sub-area. Sub-areas with zero-valued "remove" parameters are discarded next as non-informative during the training process, as demonstrated in Figure 2. Since the S-R parameters, which define the feature selection, cannot be altered during the training process, they are considered as hyperparameters. It is important to clarify that the parameters which affect the "split" and "remove" steps have no connection to the forecasting model used inside the S-R. Even if the S-R hyperparameters affect the overall performance of the forecasting model, they define the input features and not the architecture of the model, making the proposed S-R method model-agnostic.

For the BFS process, the Tree-Structured Parzen Estimator (TPE) proposed in [28], is used. TPE is a Bayesian optimisation algorithm that is commonly used for global optimisation problems in machine learning. TPE uses a probabilistic model to estimate the target function and generates the next set of candidate solutions. This Bayesian algorithm can balance exploration and exploitation, making it an effective method for finding the global optimum in high-dimensional search spaces. Additionally, TPE has been shown to perform well in problems with a noisy or non-stationary target function. Since the function that maps the S-R parameters to the error metric of a forecasting model is nearly impossible to be known, TPE approximates this mapping function using a surrogate function. Hyperparameters that minimise the surrogate function are considered as the parameters where the true mapping function is minimised. The evaluation process is explained in Section D.

The TPE algorithm, implemented within a BFS process, comprises two stages. The first stage involves modelling the prior distribution of the set of hyperparameters, denoted as h, based on the value of the loss-objective function, denoted as e. This value is obtained by training a forecasting model and is estimated by applying h to the input data from NWP. The second stage involves maximising the Expected Improvement (EI) acquisition function, which calculates the expected improvement in the loss-objective function e that would result from selecting a new candidate sample h. The calculation is performed using the previously estimated prior distribution. For the TPE algorithm to function effectively, an arbitrary number of initial sample pairs (h, e) must be generated. This is achieved by randomly sampling h and subsequently obtaining the corresponding values e by training the forecasting model for each sample. The TPE algorithm is iteratively executed by applying these two stages until a convergence criterion is satisfied.

C. Prior Distribution

In TPE, the conditional probability distribution $Pr(h|e)$ represents the likelihood of observing a set of hyperparameters h given a loss value e, which is determined by an unknown function that maps a hyperparameter vector h to the overall performance metric e of the training process. For the TPE algorithm, a threshold value $e*$ is selected, based on a quantile $\gamma \in [0, 1]$ of the observed loss values, such that $Pr(e < e^*) = \gamma$. This threshold is used to divide the observed pairs $[e, h]$ into two subsets, with densities $l(h)$ and $g(h)$, representing "valuable" and "nonvaluable" observations, respectively. Specifically, γ is chosen so that the proportion of observed loss values less than e^* is equal to γ. Smaller values of γ will result in a more conservative selection of hyperparameters, as the algorithm will only consider a smaller subset of observations as "valuable". On the other hand, larger values of γ will result in a less conservative selection of hyperparameters, as the algorithm will consider a larger subset of "valuable" observations. The choice of γ is typically problem-dependent and needs to be carefully tuned for optimal performance. In this method, the overall loss e is considered as the normalised mean absolute error obtained from the forecasting model validation process and γ is selected to be equal to 0.5. In this way, the two subsets of $[e, h]$, have equal populations. As described by (1), TPE defines $Pr(h|e)$ using these two densities.

$$Pr(h|e) = \begin{cases} l(h), & e < e^* \\ g(h), & e \geq e^* \end{cases} \tag{1}$$

Since the prior distribution $Pr(h|e)$ is based on a few sample pairs of the forecasting model's performance values e and hyperparameter sets h, it is almost impossible to do parametric density estimation, namely maximum likelihood estimation. Therefore, $l(h)$ and $g(h)$ are estimated using the Parzen window density estimation. Supposing that the values of the considered observations are defined in a D-dimensional space, then the Parzen-window density $f(h)$ is estimated as follows:

$$f(h) = \frac{1}{Nr^D} \sum_{i=1}^{N} k(\frac{h - h_i}{r}) \tag{2}$$

$$k(\frac{h - h_i}{r}) = \begin{cases} 1, & \|\frac{h-h_i}{r}\| \leq \frac{1}{2}, \quad i = 1, ..., D \\ 0, & otherwise \end{cases} \tag{3}$$

where k is the window function, h_i is the i_{th} observation and r is the length of the window function. Considering the window function as a D-dimensional hypercube with length r and given an observation h_i from N observations, the quantity $k(\frac{h-h_i}{r})$ will be equal to one if h_i falls within the hypercube centred on h with side r, and zero otherwise. Since the equation (2) is symmetrical, it can be expressed as the sum over N hypercubes, centred on the N observations h_i. Using this method, density functions $l(h)$ and $g(h)$ can be expressed using the parzen-window method, each with its own observations, as shown in equations (4) and (5). Finally, since the S-R method has multiple hyperparameters, h is defined as a vector of values and the D-dimensional space is defined over the overall number of hyperparameters, used in each TPE process.

$$l(h) = \frac{1}{Nr^D} \sum_{i=1}^{N} k(\frac{h - h_i}{r}), \quad \forall h : e < e^* \tag{4}$$

$$g(h) = \frac{1}{Nr^D} \sum_{i=1}^{N} k(\frac{h - h_i}{r}), \quad \forall h : e \geq e^* \tag{5}$$

D. EI Acquisition Function

According to [28], the parameterisation of $Pr(h, e)$ as $Pr(e)Pr(h|e)$ in the TPE algorithm is chosen to facilitate the optimisation of the expected improvement metric (EI). EI is expressed as follows:

$$EI_{e^*} = \int_{-\infty}^{\infty} max(e^* - e, 0) Pr(e|h) de \tag{6}$$

The optimization of EI under TPE is calculated as: (7).

$$\begin{aligned} EI_{e^*}(h) &= \frac{l(h)[\gamma e^* - \int_{-\infty}^{e^*} Pr(e) de]}{\gamma l(h) + (1 - \gamma) g(x)} \\ &\propto [\gamma + (1 - \gamma) \frac{g(h)}{l(h)}]^{-1} \end{aligned} \tag{7}$$

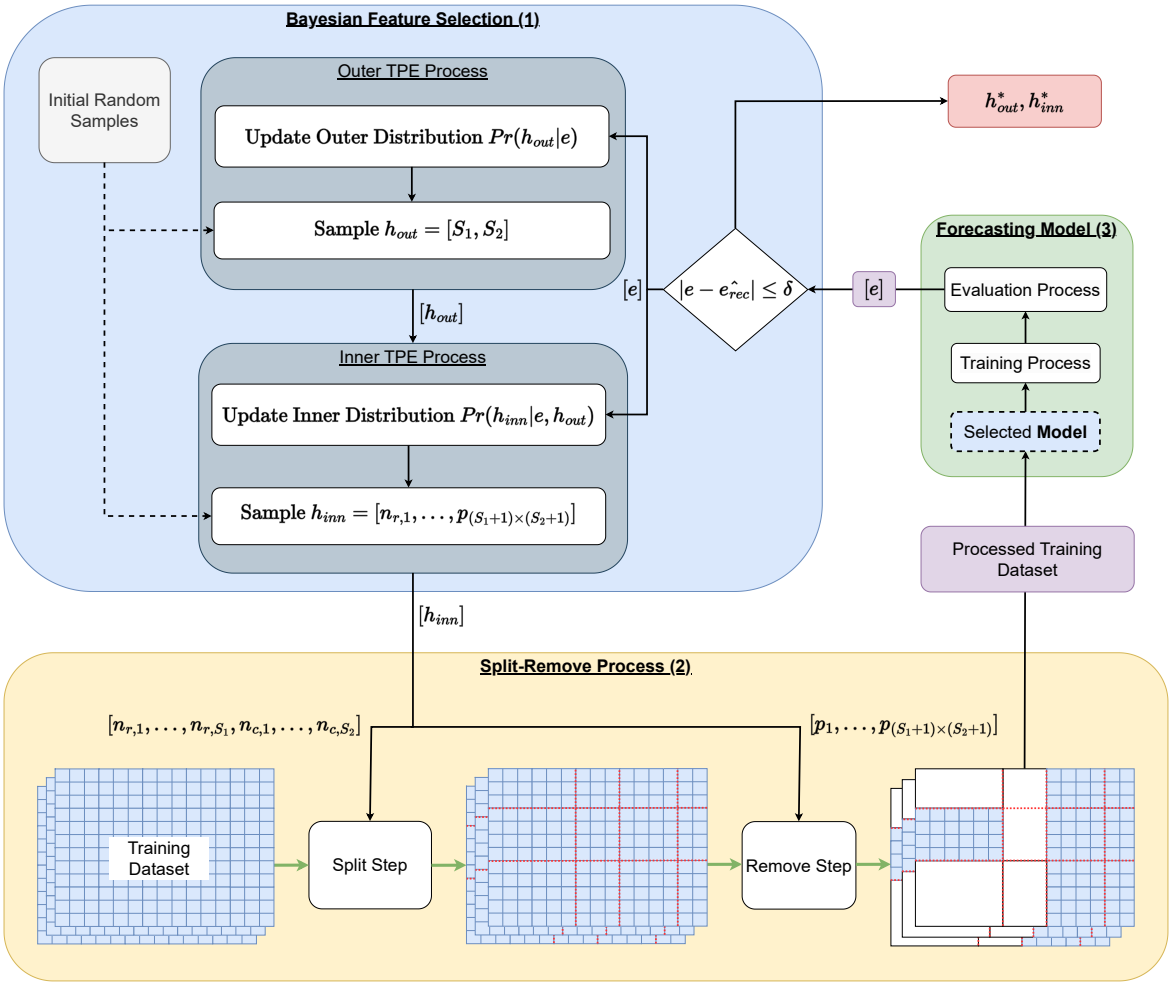


Fig. 3: Bayesian feature selection algorithm

The steps for calculating the EI under TPE are presented in more detail in Appendix I.

The expression (7) shows that to maximise the EI, the vector values of h should have a high probability under $l(h)$ and a low probability under $g(h)$. The tree-structured form of $Pr(h|e)$, makes it easy to draw samples from $l(h)$ and evaluate them according to the term $[\gamma + (1-\gamma)\frac{g(h)}{l(h)}]^{-1}$. On each iteration, the TPE algorithm chooses the candidate h^* with the highest EI.

E. S-R Steps

As defined in the previous section, the S-R process is defined by three sets of hyperparameters: The number of "split" indices in each dimension, the "split" indices of each dimension and the binary "remove" variables for each sub-area created by the "split" indices. Each set of hyperparameters has its own constraints, for the S-R process to be valid. Firstly, the number of "split" indices, in each of the first two dimensions of the input matrices, must be integer and constrained inside the range between zero and the rank of the corresponding dimension. Secondly, splitting indices must take integer values and be constrained between zero and the rank of the corresponding dimension. Finally, the "remove" hyperparameters must take binary values and at each "remove" step, at least one sub-area must be left, in order for the training process of the forecasting model to proceed. Therefore, for a NWP input matrix with spatial dimensions $d_1 \times d_2$, the S-R process requires the following hyperparameters:

- The $S_1 \in [0, d_1]$ and $S_2 \in [0, d_2]$ number of "split" indices
- The S_1 "row" splitting indices $[n_{r,1}, ..., n_{r,S_1}]$
- The S_2 "column" splitting indices $[n_{c,1}, ..., n_{c,S_2}]$
- The $(S_1 + 1) \times (S_2 + 1)$ "remove" parameters $[p_1, ..., p_{(S_1+1)\times(S_2+1)}]$

One issue that must be taken into consideration, is that hyperparameters S_1 and S_2 define the number of the rest of the hyperparameters, leading to a varying dimensionality of the D-space of possible solutions as $rank(D|S_1, S_2) = S_1 + S_2 + (S_1 + 1) \times (S_2 + 1)$. Therefore, the BFS process can be constructed as a bi-level optimization process, with two prior distributions approximated using TPE and equation (1), an outer and an inner prior distribution. A two stage TPE process is a two-step optimization procedure that leverages the results of TPE in the first stage to guide the search for the optimal solution in the second stage. In the first stage, the TPE algorithm is used to model the prior distribution of the

TABLE I: Case Studies Information

Case	Greek region	Bulgarian region	Romanian region
Input Features	U wind component (10m), V wind component (10m)	U wind component (10m), V wind component (10m)	U wind component (10m), V wind component (10m)
Forecasting Horizon (hours)	24	24	24
Time Step (hours)	1	1	1
Dataset	ENTSO-e / GFS	ENTSO-e / GFS	ENTSO-e / GFS
Training Period	01/01/2019-31/12/2019	01/01/2020-31/12/2020	01/01/2020-31/12/2020
Testing Period	01/01/2020-31/12/2020	01/01/2021-31/12/2021	01/01/2021-31/12/2021

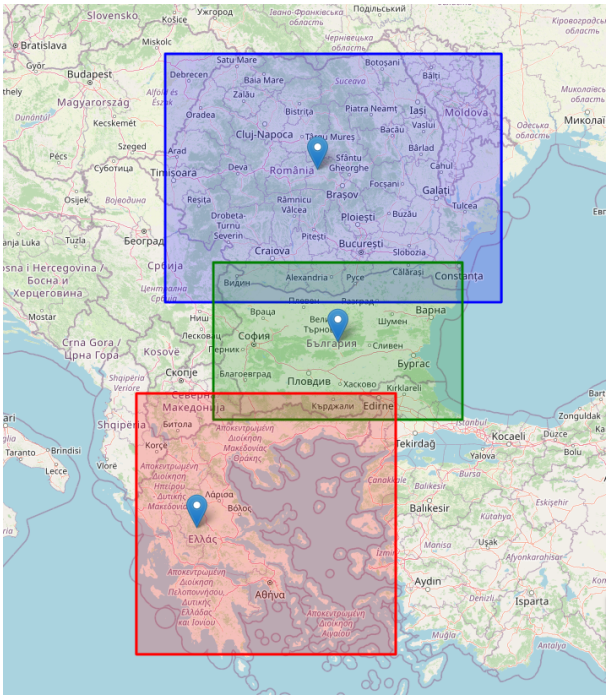


Fig. 4: Test Cases regions: Greece (Red), Bulgaria (Green) and Romania (Blue)

S_1 and S_2 hyperparameters, to maximize their EI acquisition function. The results of the first stage are used to initialize the second stage, which employs TPE to further refine the rest of the hyperparameters, conditional to the selected S_1 and S_2. Based on this approach, the hyperparameters' vector h is constructed as $h = [h_{out}, h_{inn}]$ where $h_{out} = [S_1, S_2]$ and $h_{inn} = [n_{r,1}, ..., n_{r,S_1}, n_{c,1}, ..., n_{c,S_2}, p_1, ..., p_{(S_1+1)\times(S_2+1)}]$.

The outer prior distribution models the dependency between the metric values e and $[S_1, S_2]$ hyperparameters, while the inner prior distribution models the conditional dependence between the metric values e and the rest of hyperparameters. The two prior distributions are modelled as follows:

$$Pr(h_{out}|e) = \begin{cases} l_{out}(h_{out}), & e < e^* \\ g_{out}(h_{out}), & e \geq e^* \end{cases} \quad (8)$$

$$Pr(h_{inn}|e, h_{out}) = \begin{cases} l_{inn}(h_{inn}|h_{out}), & e < e^* \\ g_{inn}(h_{inn}|h_{out}), & e \geq e^* \end{cases} \quad (9)$$

Using the distinction between the two prior distributions, the two multidimensional spaces of possible solutions, D_{inn} and D_{out}, are created. D_{out} space's rank is constant and equal to two, while the rank of D_{inn} is conditional to each sample S_1 and S_2. Furthermore, when the outer S-R process chooses an h_{out} pair that has already been investigated, the previous observed values of e and h_{inn} corresponding to that specific pair are included in the inner prior distribution, making the process more accurate and efficient.

To summarise the BFS process on the NWP, the two-stage TPE optimisation along the S-R steps is illustrated in Figure 3. In this figure, the three components of the proposed method are shown together with the type of information exchanged. In component (1), initially a number N_{out} of random samples is uniformly drawn from D_{out} and a number of N_{inn} is uniformly drawn from each D_{inn} space, corresponding to the N_{out} samples. These samples are used in the first step of the process to create the outer and inner distributions, with their corresponding error values. The outer and inner distributions are calculated based on (1) with $\gamma = 0.5$ and e^* equal to the median of the observed values of e. Once the two distributions are initialised, the iterative process follows the order of components (1), (2), and (3). At each iteration, the optimal h^*_{inn} for each h_{out} sample is calculated and then the optimal h^*_{out} is defined, based on the inner and outer EI acquisition functions, respectively. At the end of each iteration, the optimal values h^*_{out} and h^*_{inn} are integrated into the observations N_{out} and N_{inn}, in order to be included in the estimation of the two prior distributions, in the next iteration. These steps are repeated until convergence is achieved. Within this context, the absolute difference between the evaluated e and the average error of the five previous iterations $\hat{e_{rec}}$, being less than or equal to $\delta = 10^{-3}$, is assumed as a convergence criterion. It is important to note that the only information exchanged between the forecasting model and the BFS algorithm is the processed training dataset and the evaluation error that is used to update the prior distributions.

IV. Case Study

A. Evaluation Description

The proposed approach is validated on regional wind power forecasting for three countries located in south-east Europe: Greece, Bulgaria and Romania, as shown in Figure 4. Wind power measurements and installed capacity information are provided by ENTSO-e [29] and NWP data by the Global Forecasting System through the Google Earth Engine platform [30], as described in Table I. The performance of the proposed method is assessed by the bias metric, the mean absolute error (NMAE) and the mean squared error (NMSE), normalised by the installed wind power capacity of the region. The NMAE is a measure of absolute errors between paired observations and predictions, without taking into consideration the sign of

the error. The NMSE measures the squared difference between paired observations and predictions. The NMSE is the second moment of the prediction errors, which incorporates both the variance and the bias of the model. The variance is defined as the spread of the prediction errors and the bias as the distance of the predictions from the observed values. The NMSE is mostly used, along the NMAE, as it penalises the large values of prediction errors, giving a more accurate measure of the variance of the model under study. The formulas for normalised bias, NMAE and NMSE are given in (10), (11) and (12), respectively. Furthermore, bias-metric histograms are presented for a more detailed analysis of overall performance.

$$Bias(\%) = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \frac{y_i - \hat{y}_i}{IC} * 100 \tag{10}$$

$$NMAE(\%) = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \frac{|y_i - \hat{y}_i|}{IC} * 100 \tag{11}$$

$$NMSE(\%) = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\frac{y_i - \hat{y}_i}{IC})^2 * 100 \tag{12}$$

N_{obs} is the number of evaluation observations, y_i is the i_{th} observed value, $\hat{y}_i$ is the prediction of the i_{th} observation and IC is the installed capacity, all within the training process of the forecasting model.

The performance of the Bayesian feature selection method is compared with two other approaches for regional day-ahead probabilistic wind power forecasting. The first approach uses the entire area of NWP as input, while the second approach involves empirical selection of the sub-areas of NWP based on the locations of the wind farms installed in the region. The proposed S-R method is applied to three different state-of-the-art machine learning models (SVM, ANN, CNN), in order to evaluate its performance on different models. Additionally, two classical benchmark models, the Persistence model and the Climatology model, are used to provide additional performance comparisons. All models are evaluated on the same task of regional day-ahead probabilistic wind power forecasting. Models used in empirical selection of subareas are marked with a 'M' prefix and models used in the S-R method with an 'S-R' prefix. Within the validation process, the general results for all test cases are presented and to assess in more detail the performance of the proposed methodology, the Greek region is selected to investigate the seasonal evaluation metrics. Therefore, the validation period is divided into four seasons to examine the performance of the S-R method in different weather conditions in the Greek region.

B. Results and Discussion

In this section, the results for regional day-ahead wind power forecasting are presented. Specifically, the performance of the proposed Bayesian feature selection method is compared with two other approaches, as described in the previous section, and applied in three benchmark models. The performance of all models in all test cases, in terms of NMAE and NMSE, is presented in Tables II and III, respectively. It is clear

TABLE II: NMAE evaluation results for each test case

NMAE (%)			
Model	Greece	Bulgaria	Romania
Persistence	12.553	20.489	23.488
Climatology	13.586	15.176	19.811
SVM	6.089	5.992	9.652
M-SVM	5.571	5.687	9.169
S-R SVM	5.154	5.452	9.315
ANN	7.765	5.321	7.855
M-ANN	6.433	4.981	7.305
S-R ANN	5.803	4.590	7.149
CNN	5.951	5.234	10.375
M-CNN	5.506	4.678	9.856
S-R CNN	4.852	4.324	9.166

TABLE III: NMSE evaluation results for each test case

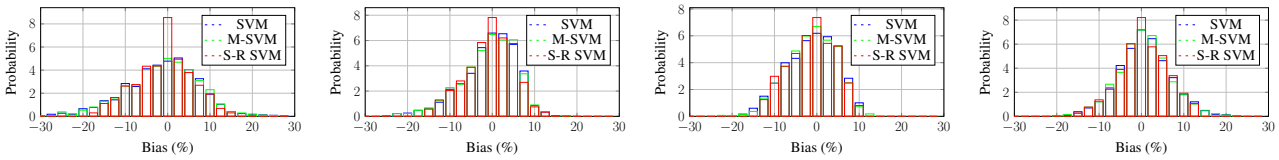
NMSE (%)			
Model	Greece	Bulgaria	Romania
Persistence	2.7695	5.128	11.924
Climatology	2.647	3.557	5.886
SVM	0.489	0.513	1.472
M-SVM	0.471	0.449	1.383
S-R SVM	0.436	0.415	1.314
ANN	1.118	1.062	1.284
M-ANN	0.775	0.736	1.212
S-R ANN	0.605	0.574	1.091
CNN	0.642	0.603	2.004
M-CNN	0.553	0.506	1.67
S-R CNN	0.489	0.464	1.386

TABLE IV: Seasonal NMAE evaluation results for test case: Greece

NMAE (%)				
Model	Winter	Spring	Summer	Autumn
Persistence	16.745	12.166	10.392	10.909
Climatology	15.596	12.496	13.814	12.438
SVM	7.769	6.163	5.386	5.038
M-SVM	7.061	5.521	4.724	4.974
S-R SVM	6.661	5.162	4.441	4.350
ANN	11.606	8.514	5.264	5.679
M-ANN	9.117	6.888	4.487	5.241
S-R ANN	8.167	6.107	4.045	4.894
CNN	7.778	5.651	4.820	5.558
M-CNN	6.975	5.447	4.546	5.057
S-R CNN	6.007	4.942	4.003	4.456

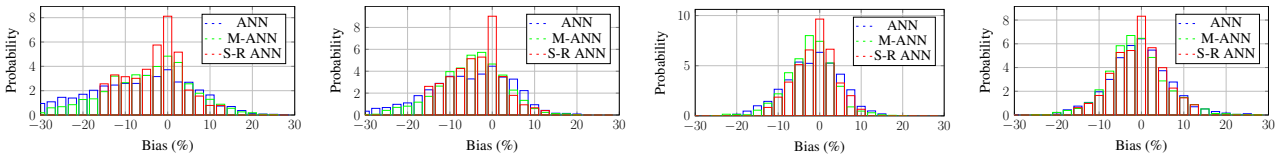
TABLE V: Seasonal NMSE evaluation results for test case: Greece

NMSE (%)				
Model	Winter	Spring	Summer	Autumn
Persistence	4.507	2.589	1.915	2.067
Climatology	3.528	2.326	2.656	2.079
SVM	0.744	0.443	0.354	0.418
M-SVM	0.722	0.414	0.335	0.415
S-R SVM	0.689	0.369	0.301	0.385
ANN	2.302	1.189	0.433	0.549
M-ANN	1.488	0.789	0.347	0.486
S-R ANN	1.067	0.622	0.309	0.424
CNN	1.017	0.554	0.434	0.561
M-CNN	0.862	0.501	0.337	0.515
S-R CNN	0.771	0.427	0.295	0.462



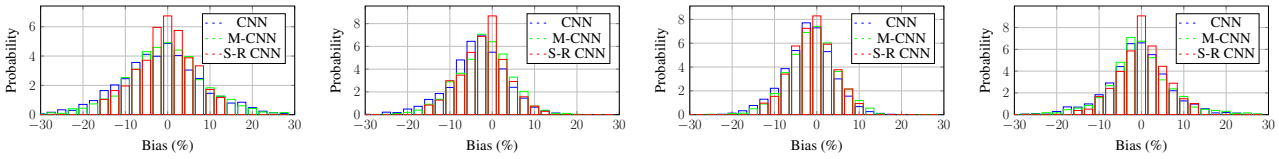
(a) (b) (c) (d)

Fig. 5: Histogram of bias in SVM results for (a): Winter (b): Spring (c): Summer (d): Autumn



(a) (b) (c) (d)

Fig. 6: Histogram of bias in ANN results for (a): Winter (b): Spring (c): Summer (d): Autumn



(a) (b) (c) (d)

Fig. 7: Histogram of bias in CNN results for (a): Winter (b): Spring (c): Summer (d): Autumn

that the proposed method outperforms the use of the whole NWP area and the empirical selection of sub-areas along with the rest of the other feature selection methods in all three machine learning models. An important conclusion is that the dimensionality reduction achieved through the selection of sub-areas, either empirically or by the Bayesian method, seems to improve the overall performance of the forecasting models. Performance improvement expressed by both in NMAE and NMSE is achieved in the three benchmark models, while the variance of the prediction errors is lower. Regarding the NMAE and NMSE metrics, as observed in Tables II and III, the robustness of the proposed methodology is proven, as in the three test cases, the use of the S-R method, succeeds in improving the accuracy of the prediction and decreasing the deviations of forecast errors, regardless the regions complexity of terrain. One major point of the results of the proposed method is that the performance improvements are achieved by significantly decreasing the number of input features, through the removal of NWP sub-areas. In the case of Greece, the initial number of input features is 1334 and the final number of input features equal to 756, which leads to a percentage decrease of input features equal to 43%. Similarly, in the test case of Bulgaria, the percentage decrease in input features is equal to 56% ($784 \rightarrow 344$) and in the test case of Romania equal to 30% ($1330 \rightarrow 948$).

In order to further investigate the performance of the proposed method, a detailed evaluation of its seasonal performance is presented for the test case of Greece. In addition to the previous results, the NMAE and NMSE metrics for each season are shown in Tables IV and V. To further this investigation, the bias histograms of the different models are presented in Figures 5-7. What can be observed from Tables IV and V, is that the performance of all models varies in different seasons, due to the seasonal variability of weather conditions and wind cycles. However, the S-R method can improve the accuracy of the forecasting models it is applied to, regardless of the different weather conditions. This can also be observed in the bias histograms produced by applying the S-R method, compared to the rest of the distributions, indicating lower width or smaller deviation from the expected value. Additionally, the expected bias of error in each case is decreased by the S-R method, indicating that the model is more unbiased in over-predicting or under-predicting.

To illustrate the performance of the different approaches in the selection of subareas, the results obtained by three selection methods are presented in Figures 8a-8c. Figure 8a displays the entire NWP over the region, without any sub-area selection. The red crosses indicate the installed wind farms. In Figure 8b, the sub-areas are empirically selected based on the locations of the wind farms. As can be seen, these subareas are scattered throughout the region and do not form a clear pattern. Finally, in Figure 8c, the selection of sub-areas obtained through the S-R CNN method is displayed. In this case, the selected sub-areas create a pattern, where sub-areas over some existing wind farms are removed, since they do not contribute to the improvement of the accuracy of the forecasting models. An important observation is that the sub-areas include areas over the sea and the coastline of the country, while parts of the country where the terrain is irregular and contains mountainous regions are avoided, even

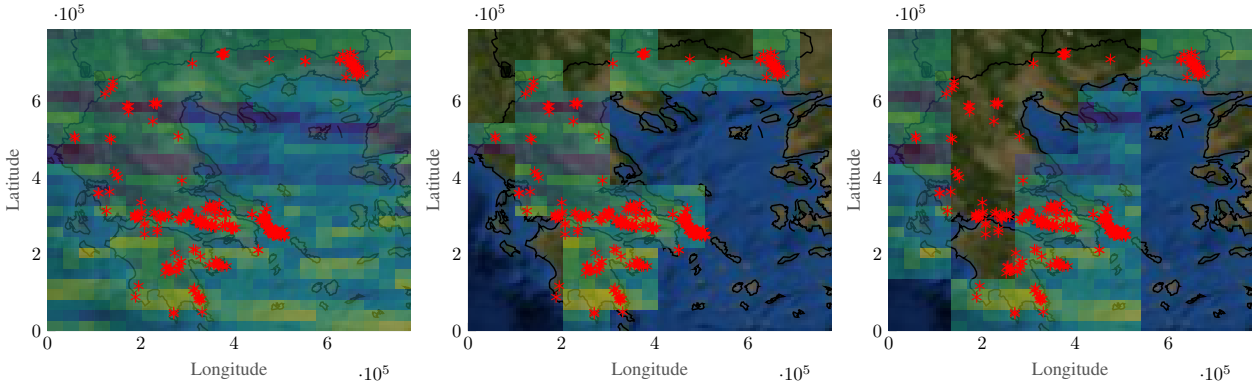


Fig. 8: (a): NWP over the entire region (b): Empirical selection of sub-areas over existing wind farms (c): S-R based selection of sub-areas over the region

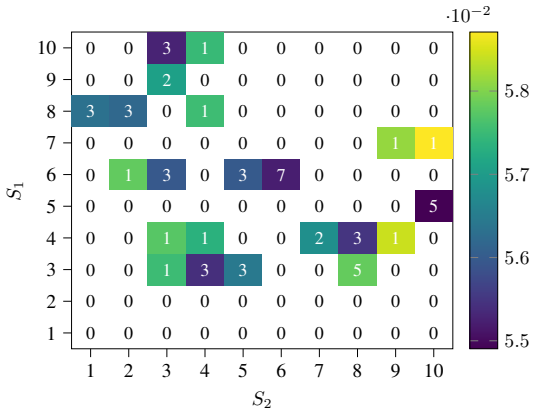


Fig. 9: Split numbers' values through the optimisation process

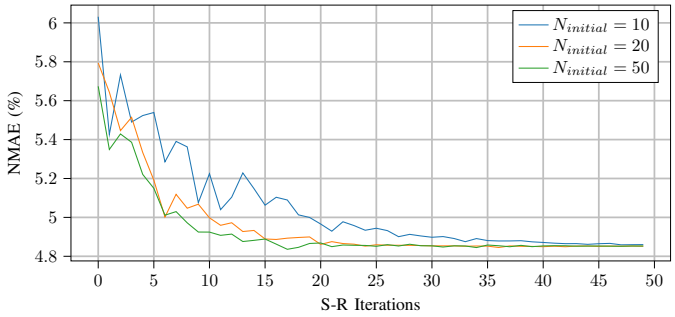


Fig. 10: Sensitivity analysis of the S-R method to the number of initial random samples

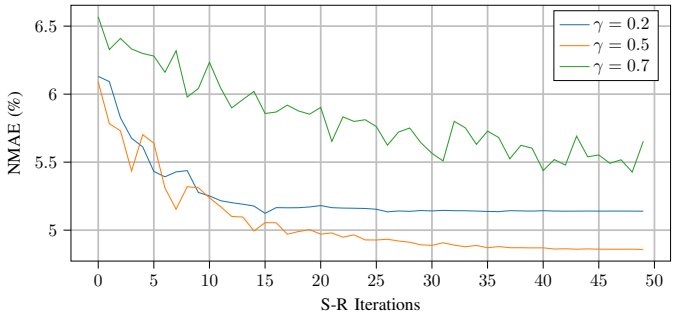


Fig. 11: Sensitivity analysis of the S-R method to the γ parameter

if they contain installed wind farms. This might suggest that even if some areas contain wind farms, the accuracy obtained by using the NWP for these areas is poor, resulting in a reduced performance of the forecasting model. This is the main advantage of the S-R method, which is based on the actual performance of the forecasting model used and the actual measured wind power generation of the area under study.

Furthermore, in order to validate the Bayesian behaviour of the S-R method, using a two-stage TPE algorithm, the results of the outer optimisation process for the S-R CNN are presented in Figure 9. Due to visualisation limitations caused by the higher dimensionality of the inner optimisation process, only the results of the outer process are displayed, in a matrix form, where each cell includes the number of iterations each pair $[S_1, S_2]$ was drawn randomly from the prior distribution at each step. The colour of each cell corresponds to the minimum NMAE, achieved within the inner optimisation process of the corresponding pair $[S_1, S_2]$, as defined by the matrix indices. It can be observed that during the optimisation process, areas with high and low probability of good performance are created in the outer prior distribution, as it is updated at each step. S_1 and S_2 pairs with good performance have been selected multiple times, due to their high probability of resulting lower NMAE values, while other pairs have been selected only a few times or even not at all. This proves the effectiveness of the S-R method, which is guided by the previous results of the S-R process and the training of the forecasting model on which it is applied.

Finally, the sensitivity of the proposed method to its various parameters is studied. In Figures 10 and 11, the convergence trajectories for the S-R CNN model for the Greek test case are shown, for three different numbers of initial random samples and for three different values γ, respectively. The initial random samples define the solution space for both the outer and inner optimisation steps. Variation in the number of

initial samples, as displayed, affects the speed of convergence of the algorithm, where a larger number of initial samples leads to a faster convergence. Furthermore, as proven by the formulas of the TPE ((4),(5)), the γ value acts as a trade-off between the exploration and the exploitation of the solution space. The three trajectories in Figure 11, illustrate this trade-off, as higher values of γ lead to greater exploration of the solution space with a lower probability of convergence, due to the high oscillations of the values of e. On the other hand, lower values of γ lead to greater exploitation of the solution space, achieving convergence in fewer iterations, but they have the potential to be trapped in a local minimum. Therefore, a value of γ equal to 0.5, is a good candidate for balanced exploration and exploitation of the TPE algorithm.

V. Conclusion

This paper shows that dimensionality reduction of NWP inputs with BFS, through elimination of sub-areas, can improve the accuracy of regional day ahead wind power forecasting. The proposed method using a two-stage TPE process proved to be more effective than using the entire NWP area or an empirical selection of subareas. Moreover, as the proposed method is applied directly to the input data, it can be used with any forecasting model, making it model-agnostic and widely applicable. The proposed method is applied for country-wide day-ahead wind power forecasting in three countries in south-east Europe and has been proven to increase the performance of popular data-driven methods applied for this task.

References

[1] R. Tawn and J. Browell, "A review of very short-term wind and solar power forecasting," *Renewable and Sustainable Energy Reviews*, vol. 153, p. 111758, 2022.

[2] X. Wang, P. Guo, and X. Huang, "A review of wind power forecasting models," *Energy Procedia*, vol. 12, pp. 770–778, 2011, the Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011.

[3] G. Giebel and G. Kariniotakis, "Wind power forecasting—a review of the state of the art," *Renewable energy forecasting*, pp. 59–109, 2017.

[4] G. Sideratos and N. D. Hatziargyriou, "A distributed memory rbf-based model for variable generation forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 106041, 2020.

[5] J. Henze, M. Siefert, S. Bremicker-Trübelhorn, N. Asanalieva, and B. Sick, "Probabilistic upscaling and aggregation of wind power forecasts," *Energy, Sustainability and Society*, vol. 10, 03 2020.

[6] W. Dong, H. Sun, J. Tan, Z. Li, J. Zhang, and H. Yang, "Regional wind power probabilistic forecasting based on an improved kernel density estimation, regular vine copulas, and ensemble learning," *Energy*, vol. 238, p. 122045, 2022.

[7] H. Zhang, Y. Liu, J. Yan, S. Han, L. Li, and Q. Long, "Improved deep mixture density network for regional wind power probabilistic forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2549–2560, 2020.

[8] Z. Wang, W. Wang, C. Liu, and B. Wang, "Forecasted scenarios of regional wind farms based on regular vine copulas," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 1, pp. 77–85, 2020.

[9] X. Dong, Y. Sun, Y. Li, X. Wang, and T. Pu, "Spatio-temporal convolutional network based power forecasting of multiple wind farms," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 2, pp. 388–398, 2022.

[10] Z. Wang, W. Wang, C. Liu, B. Wang, and S. Feng, "Short-term probabilistic forecasting for regional wind power using distance-weighted kernel density estimation," *IET Renewable Power Generation*, vol. 12, 09 2018.

[11] H. Apaydin, F. K. Sonmez, and Y. E. Yildirim, "Spatial interpolation techniques for climate data in the gap region in turkey," *Climate Research*, vol. 28, no. 1, pp. 31–40, 2004.

[12] M. Cellura, G. Cirrincione, A. Marvuglia, and A. Miraoui, "Wind speed spatial estimation for energy planning in sicily: A neural kriging application," *Renewable energy*, vol. 33, no. 6, pp. 1251–1266, 2008.

[13] M. B. Ozkan and P. Karagoz, "Data mining-based upscaling approach for regional wind power forecasting: Regional statistical hybrid wind power forecast technique (regionalshwip)," *IEEE Access*, vol. 7, pp. 171 790–171 800, 2019.

[14] P. Pinson, N. Siebert, and G. Kariniotakis, "Forecasting of regional wind generation by a dynamic fuzzy-neural networks based upscaling approach," *Proceedings EWEC 2003 (European Wind energy and conference)*, 06 2003.

[15] S. Basu, S. Watson, E. Lacoa Arends, and B. Cheneka, "Day-ahead wind power predictions at regional scales: Post-processing operational weather forecasts with a hybrid neural network," 10 2020.

[16] W. Dong, H. Sun, J. Tan, Z. Li, J. Zhang, and Y. Y. Zhao, "Short-term regional wind power forecasting for small datasets with input data correction, hybrid neural network, and error analysis," *Energy Reports*, vol. 7, pp. 7675–7692, 2021.

[17] D. A. Wood, "Trend decomposition aids short-term countrywide wind capacity factor forecasting with machine and deep learning methods," *Energy Conversion and Management*, vol. 253, p. 115189, 2022.

[18] Y. Yu, M. Yang, X. Han, Y. Zhang, and P. Ye, "A regional wind power probabilistic forecast method based on deep quantile regression," *IEEE Transactions on Industry Applications*, vol. 57, no. 5, pp. 4420–4427, 2021.

[19] T. Ding, P. Li, G. Huang, Y. Yu, Z. Si, F. Yan, X. Liu, and M. Li, "A statistical upscaling approach of region wind power forecasting based on combination model," in *2020 IEEE 3rd Student Conference on Electrical Machines and Systems (SCEMS)*, 2020, pp. 596–601.

[20] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geoscience and remote sensing letters*, vol. 4, no. 4, pp. 674–677, 2007.

[21] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1360–1367, 2001.

[22] E.-S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid, and S. E. Hussein, "Novel feature selection and voting classifier algorithms for covid-19 classification in ct images," *IEEE Access*, vol. 8, pp. 179 317–179 335, 2020.

[23] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[24] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, 2022.

[25] Y. Yang, W. Li, T. A. Gulliver, and S. Li, "Bayesian deep learning-based probabilistic load forecasting in smart grids," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4703–4713, 2020.

[26] Y. Chen, Y. Wang, P. Ren, M. Wang, and M. de Rijke, "Bayesian feature interaction selection for factorization machines," *Artificial Intelligence*, vol. 302, p. 103589, 2022.

[27] O. Abedinia, N. Amjady, and H. Zareipour, "A new feature selection technique for load and price forecast of electrical power systems," *IEEE Transactions on Power Systems*, vol. 32, no. 1, pp. 62–74, 2017.

[28] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[29] European Network of Transmission System Operators for Electricity (ENTSO-E), "ENTSO-E Transparency Platform," 2023. [Online]. Available: https://transparency.entsoe.eu/

[30] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017.

Appendix I: Calculation of Expected Improvement under Tree-Structured Parzen Estimator

In this appendix, we provide a detailed explanation of how the EI acquisition function is calculated under the TPE approach. TPE as a Bayesian optimization technique, constructs a probabilistic model of the objective function based on observations of the function evaluations at different hyperparameter

settings. The model consists of two parts: a prior probability distribution over the hyperparameters and a likelihood function that models the relationship between the hyperparameters and the objective function values. TPE uses these two parts to construct an acquisition function that guides the search towards promising regions of the hyperparameter space.

The EI acquisition function is defined as follows:

$$EI = \int_{-\infty}^{\infty} max(e^* - e, 0) Pr(e|h) de \tag{13}$$

$$max(e^* - e, 0) = \begin{cases} e^* - e, & e < e^* \\ 0, & e \geq e^* \end{cases} \tag{14}$$

Therefore, equation (13) is simplified into the following equation:

$$EI = \int_{-\infty}^{e^*} (e^* - e) Pr(e|h) de \tag{15}$$

Based on Bayes' theorem, the conditional probability of error e in respect to hyperparameters set h can be described as follows:

$$Pr(e|h) = \frac{Pr(h|e)Pr(e)}{Pr(h)}, \quad Pr(h) \neq 0 \tag{16}$$

By construction, as defined in (1), $Pr(h|e)$ is:

$$Pr(h|e) = \begin{cases} l(h), & e < e^* \\ g(h), & e \geq e^* \end{cases} \tag{17}$$

The term $Pr(e < e^*) = \gamma$ and $Pr(h)$ is defined as follows:

$$Pr(h) = \int_R Pr(h|e)Pr(e)de = \gamma l(h) + (1 - \gamma) g(h) \tag{18}$$

Using the above equations, (15) is transformed into:

$$EI = \int_{-\infty}^{e^*} (e^* - e) \frac{Pr(h|e)Pr(e)}{Pr(h)} de \tag{19}$$

$$EI = \int_{-\infty}^{e^*} (e^* - e) \frac{l(h)Pr(e)}{\gamma l(h) + (1 - \gamma) g(h)} de \tag{20}$$

$$EI = \frac{l(h)}{\gamma l(h) + (1 - \gamma) g(h)} \int_{-\infty}^{e^*} (e^* - e) Pr(e) de \tag{21}$$

As it can be observed, the term $\int_{-\infty}^{e^*} (e^* - e) Pr(e) de$ is independent of the hyperparameters set h, therefore it can be ignored in the optimization process of EI. Finally, the EI that has to be maximized in respect to h is defined as follows:

$$EI \propto \frac{l(h)}{\gamma l(h) + (1 - \gamma) g(h)} = [\gamma + (1 - \gamma) \frac{g(h)}{l(h)}]^{-1} \tag{22}$$