Better Wind forecasting using Evolutionary Neural Architecture Search driven Green Deep Learning

Keerthi Nagasree Pujari [a,c,#], Srinivas Soumitri Miriyala[a,d,#], Prateek Mittal[b,e] & Kishalay Mitra[a,*]

[a]Department of Chemical Engineering, Indian Institute of Technology Hyderabad, India

[b]Department of Computer Science, University College London, UK

[c]ch20resch11006@iith.ac.in, [d]srinivas.soumitri@gmail.com, [c]mittal.prat@gmail.com

[#]Equal contribution

[*]Corresponding Author

Declaration of interests: None

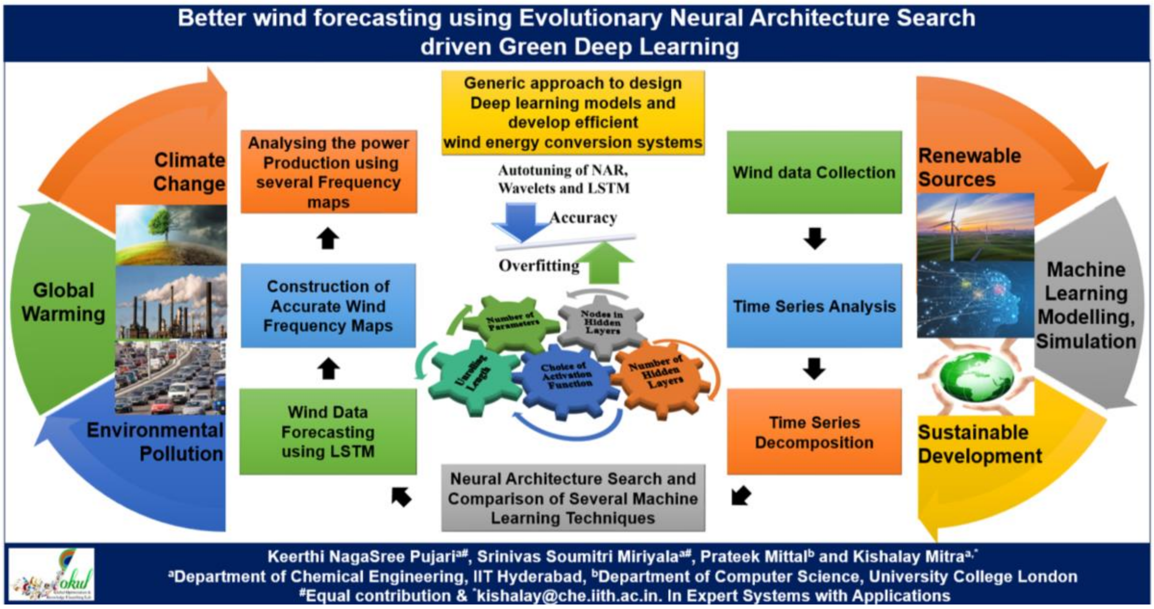
[*]Corresponding Author: kishalay@che.iith.ac.in, Department of Chemical Engineering, Indian Institute of Technology Hyderabad, Kandi, Telangana 502284, India.

ABSTRACT

Climate Change heavily impacts global cities, the downsides of which can be minimized by adopting renewables like wind energy. Despite its advantages, the nonlinear nature of wind renders the forecasting approaches to design and control wind farms ineffective. To expand the research horizon, the current study a) analyses and performs decomposition of real-world time-series wind data, b) identifies the modelling issues associated with state-of-the-art methods such as Nonlinear Auto-Regressive (NAR) models, Wavelet Neural Networks (WNN), and Long Short-term Memory (LSTM) networks, and c) presents a comprehensive comparison between them for accurate wind forecasting. The present work also focuses on another important aspect related to the design of the aforementioned networks. This work proposes an evolutionary strategy for Neural Architecture Search (NAS) with the objective to minimize the computational cost associated with training and inferring the networks which form the central theme of Green Deep Learning. Balancing the trade-off between parsimony and prediction accuracy, the proposed NAS strategy could optimally design NAR, WNN, and LSTM models with a mean test accuracy of 99%. The robust methodologies discussed in this work not only accurately model the wind behaviour but also provide a green generic approach for designing deep neural networks.

Keywords: Renewable Energy; Wind Characteristics forecasting; Neural Architecture Search; Green Deep Learning; Effective Wind Farm Design.

Graphical Abstract



Better wind forecasting using Evolutionary Neural Architecture Search driven Green Deep Learning

Generic approach to design Deep learning models and develop efficient wind energy conversion systems

Analysing the power Production using several Frequency maps

Construction of Accurate Wind Frequency Maps

Wind Data Forecasting using LSTM

Wind data Collection

Time Series Analysis

Time Series Decomposition

Neural Architecture Search and Comparison of Several Machine Learning Techniques

Keerthi NagaSree Pujari[a#], Srinivas Soumitri Miriyala[a#], Prateek Mittal[b] and Kishalay Mitra[a,*]
[a]Department of Chemical Engineering, IIT Hyderabad, [b]Department of Computer Science, University College London
[#]Equal contribution & [*]kishalay@che.iith.ac.in. In Expert Systems with Applications

Abbreviations

ADAM	Adaptive Momentum
ADF	Augmented-Dickey-Fuller
AEP	Annual Energy Production
AIC	Akaike Information Criterion
autoML	automated Machine Learning
BDS	Brock-Dechert-Scheinkman
INLP	Integer Nonlinear Programming
LSTM	Long Short Term Memory Networks
NAR	Nonlinear Autoregressive Models
NAS	Neural Architecture Strategy
NSGA-II	Non-dominated Sorting Genetic Algorithm II
NWP	Numerical Weather Prediction
PMF	Probability Mass Function
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
STL	Seasonal and Trend decomposition using Loess
t-BPTT	truncated-Backpropagation Through Time
WFM	Wind Frequency Map
WNN	Wavelet Neural Networks

Nomenclature:

$\tilde{C}_i^{m,p}$	Intermittent Cell value of i^{th} node in m^{th} hidden layer at time step p
$\hat{X}^t$	Estimated data at time step t
b_i^m	bias of i^{th} node in m^{th} layer
B_{LB}^T, B_{UB}^T	Lower and upper bounds on B^T

B^{T}	Length of Unrolled Network
$\mathrm{C}_{\mathrm{i}}^{\mathrm{m,p}}$	Cell state in ith node in mth hidden layer at time step p
$\mathrm{F}_{\mathrm{i}}^{\mathrm{m,p}}$	Forget gate of ith node in mth hidden layer at time step p
$\mathrm{I}_{\mathrm{i}}^{\mathrm{m,p}}$	Input gate of ith node in mth hidden layer at time step p
$\mathrm{M}_{\mathrm{LB}}, \mathrm{M}_{\mathrm{UB}}$	Lower and upper bound on M
$\mathrm{N}_{\mathrm{LB}}, \mathrm{N}_{\mathrm{UB}}$	Lower and upper bounds on N^{m}
N^{m}	number of nodes in m^{th} layer
N_{P}	Number of parameters in the model
$\mathrm{O}_{\mathrm{i}}^{\mathrm{m,p}}$	Output gate of ith node in mth hidden layer at time step p
$\mathrm{P}_{\mathrm{curve}}$	power curve
$\overline{\mathrm{T}}$	Number of Test data points
T^{F}	Forward propagation length in t-BPTT
T_{ij}	number of points in i^{th} direction sector and j^{th} speed bin
$\mathrm{u}_{\mathrm{effective}}$	effective velocity at a given turbine obtained after application of wake
u_{r}	values of speed in rth interval
$\mathrm{w}_{\mathrm{ij}}^{\mathrm{m}}$	weight on connection from j^{th} node in $(\mathrm{m}-1)^{\mathrm{th}}$ layer to i^{th} node in m^{th} layer
$\mathrm{x}_{\mathrm{i}}^{\mathrm{m}}$	activated output of i^{th} node in m^{th} hidden layer
X^{t}	data at time step t
$\mathrm{z}_{\mathrm{ij}}^{\mathrm{m}}$	Translated and dilated variable in m^{th} hidden layer from j^{th} node to i^{th} node in a WNN
$y_{\mathrm{i}}^{\mathrm{m}}$	weighted sum of inputs
A	Activation function in LSTMs
D	Number of Direction sectors

H	Hurst exponent
K	dimension of data
L	Loss function
M	Number of layers in the network (Hidden layers + output layer)
R^2	Correlation coefficient
T	Length of data
U	Number of Speed bins

Greek Symbol:

f	functional map
θ	Parameters of Neural network
$\mathcal{F}_{ij}$	frequency in ith direction sector and jth speed bin
ϕ_q	values of direction in qth interval
$\mathbb{N}$	Number of Turbines
φ	activation function in NAR and Wavelets
Ψ	output after application of a wavelet transform on z

Symbol:

GW	Gigawatt
MW	Megawatt
CO_2	Carbon dioxide
sq. Km	Square kilometre

1. Introduction

The rapid growth of human civilization has led to a 67% surge in energy demand across the world in the past three decades(*World Energy Consumption Statistics | Enerdata*, 2021). According to the Global Energy Yearbook 2021, fossils fuels account for ~81% of the total energy consumption resulting in a 24% increase in CO_2 emissions(*World Energy Consumption Statistics | Enerdata*, 2021). Numerous conferences on climate change starting from the Earth Summit in 1992 (Grubb et al., 2019) to the 26th United Nations Climate change conference of the Parties (COP26) in Glasgow (Vogler, 2021) brought the world leaders together to address the issue of global warming and climate change by mitigation of greenhouse-gas-emissions. As a result of these efforts, the utilization of renewable energy sources has steadily increased

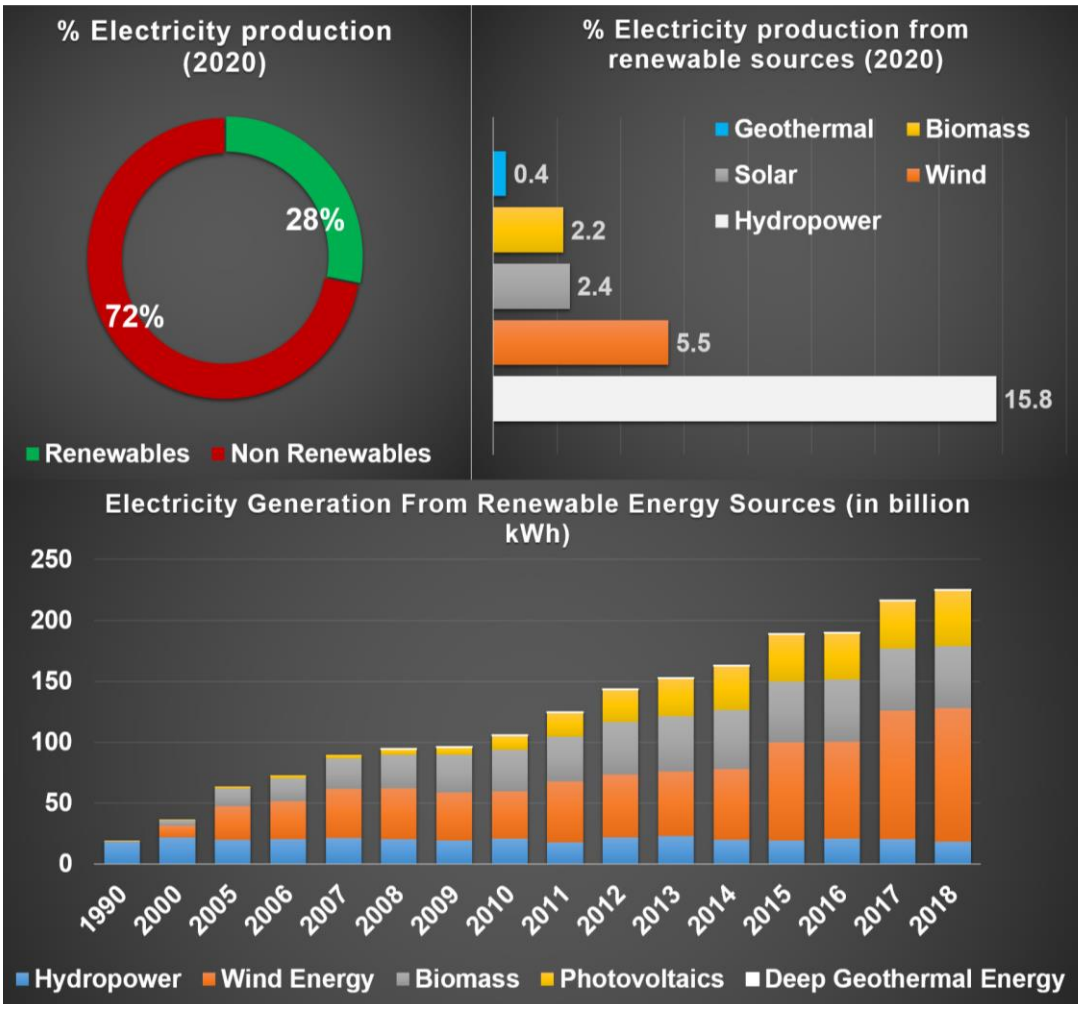


Figure 1. Trends in utilization of renewable sources (1990-2020) (Enerdata, 2021)

in the last three decades as shown in Figure 1 (*Renewables in Electricity Production | Statistics Map by Region | Enerdata*, 2021).

Among various alternative sources of renewable energy generation wind has attracted significant attention from researchers, practitioners, investors, and policymakers due to the aspects of easy availability, cleaner production, and scope for large-scale generation. As per the Global Wind Report 2021, the total cumulative installations of wind energy have reached 743 GW helping to avoid over 1.1 billion tonnes of CO_2 globally(*Global Wind Report 2021 - Global Wind Energy Council*, 2021). With 95 GW installations in 2020 alone (~53% year-on-year increase), wind farms have emerged as the clean energy technology with the most decarbonization potential per MW. However, the report suggests that this rate needs to be tripled in the coming decade to stay on the path towards net carbon neutrality by 2050, calling for urgent action from policymakers to scale up the wind power production at the necessary pace (*Global Wind Report 2021 - Global Wind Energy Council*, 2021).

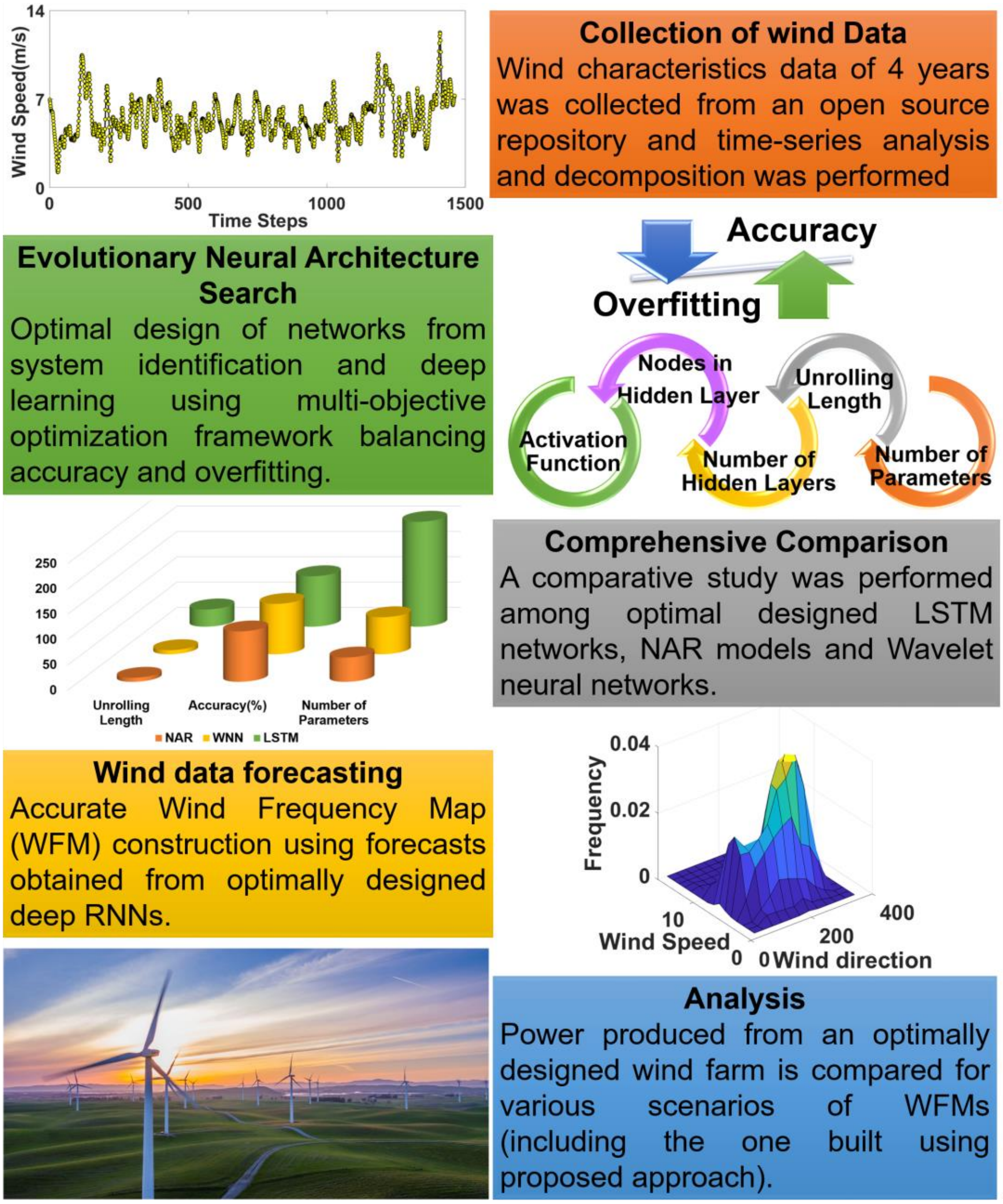
Despite so much focus on wind energy, one of the biggest challenges it faces is its uncertain nature, which results in tremendous variability in energy production. Therefore, accurate prediction of its variability can be of great help to the wind-farm owners and the industries in planning and execution of better energy conversion and management systems. Conventionally, the wind is modeled by constructing a Probability Mass Function (PMF) using time-series data of wind speed and direction. This PMF is then used in applications such as wind-farm layout optimization (or micro-siting) and control (Ciri et al., 2019; Miao et al., 2018). Though this method in practice is the best possible practical way to handle the variability in wind, since the PMF is built on a limited amount of time-series data, the ability to capture long-range variability in wind is sacrificed, making the results unrealistic. This necessitates the requirement of novel methods capable of forecasting accurately by considering the long-term variability in the data. The importance of forecasting in the wind energy domain is presented in many recent articles (Allen et al., 2017; Chen et al., 2022; L. L. Li et al., 2021; X. Liu et al., 2021; Z. Liu et al., 2021; Song et al., 2011; Jianzhou Wang et al., 2021; K. Wang et al., 2018; Xue et al., 2020).

When it comes to modeling nonlinear trends in wind characteristics, traditionally, physics-based methods (e.g. Numerical Weather Prediction (NWP)) have been employed. The complexity in modeling weather conditions using first principles, lack of professional staff for collection and maintenance of crucial data to validate these models, and high computational costs make these physics-based models difficult to handle. These difficulties of physics-based models turned researchers towards data-driven techniques. Under this category, researchers are found to be inclined towards the utilization of system identification tools such as linear and Nonlinear Auto-Regressive (NAR) models, Fuzzy inference systems, and Wavelet Neural Networks (WNNs) for modeling and forecasting wind characteristics (Abhinav et al., 2017; An et al., 2011; Brahimi, 2019; Daniel et al., 2020; Jahangir et al., 2020; Prasetyowati et al., 2017; Salcedo-Sanz et al., 2011; Zhang et al., 2022). Apart from conventional system identification techniques, the applicability of deep learning (such as Long Short Term Memory Networks (LSTMs), and Gated Recurrent Units) has been increasing in recent times due to its ability to handle extreme transience and nonlinearities in data such as that in the wind time series (Ding et al., 2019; Kumar Dubey et al., 2021; H. Li et al., 2022; Ningsih et al., 2019; Olaofe, 2014; Trebing & Mehrkanoon, 2020; Jujie Wang et al., 2022; Y. X. Wu et al., 2019).

The ease in availability of open-source software for system identification and machine learning techniques helped in the tremendous rise of their applicability. However, in open source software, the difficulty arises with the selection of hyperparameters, e.g. topology of the network, choice of activation functions, etc., which govern the predictability of these models. To overcome the difficulty in heuristics associated with machine learning models, many recent works reported reinforcement learning, Bayesian optimization, and single-objective optimization-based frameworks (Cho et al., 2020; Dong et al., 2021; Han et al., 2020; Tso et al., 2020; J. Wu et al., 2020, 2019). To the best of our knowledge, no work has been reported, that discusses the optimal design of state-of-the-art nonlinear system identification tools as well as deep learning methods, using a single, holistic, multi-objective evolutionary optimization-based framework balancing the aspects of over-fitting and accuracy and comparing them in terms of modeling long-term variability in wind data and predictability.

1.1 Contributions

The aim of the current study is, therefore, to develop optimal methodologies for modeling the real, nonlinear, transient wind characteristics data. To achieve this (see Figure 2), first the nature of data (nonlinearity, stationarity, and long-term dependency), is examined to justify the application of appropriate time-series modeling techniques. The hidden patterns and the effect of periodicities are then studied using STL decomposition. It is then proposed to use state-of-the-art modeling techniques from nonlinear system identification (NAR and WNNs) and deep learning (LSTMs). In all these techniques, the model hyperparameters (e.g. number of hidden layers, number of nodes in each hidden layer, choice of activation function, and number of unrolled time steps or the order of the model), which are conventionally fixed using heuristics thereby providing scope for inefficiencies, are estimated intelligently using optimal evolutionary search. The two conflicting attributes of data-based modeling: maximization of model accuracy and minimization of model complexity drive the evolutionary neural architecture search strategy proposed in this work. Minimization of the model complexity reduces the computations required by the optimally designed models thereby significantly decreasing the associated carbon footprint (Xu et al., 2021). The smaller optimally designed models have a high rate of deployment in real-world applications. The proposed methodology thus contributes to Green Deep Learning (Xu et al., 2021).After successfully training, the credibility of the forecasts from optimally designed models is validated by comparing with



Collection of wind Data

Wind characteristics data of 4 years was collected from an open source repository and time-series analysis and decomposition was performed

Evolutionary Neural Architecture Search

Optimal design of networks from system identification and deep learning using multi-objective optimization framework balancing accuracy and overfitting.

Comprehensive Comparison

A comparative study was performed among optimal designed LSTM networks, NAR models and Wavelet neural networks.

Wind data forecasting

Accurate Wind Frequency Map (WFM) construction using forecasts obtained from optimally designed deep RNNs.

Analysis

Power produced from an optimally designed wind farm is compared for various scenarios of WFMs (including the one built using proposed approach).

Figure 2. Overall framework of wind characteristics forecasting and its application proposed in this manuscript.

realistic wind characteristic data collected over 4 years from a wind farm in France. Additionally, a comparative study is performed among the optimal NAR, optimal WNNs, and optimal LSTMs by demonstrating the applicability of forecasts over a long range of time in the optimal design of a wind energy conversion system.

In the rest of the paper, Section 2 presents time-series decomposition and analysis techniques in brief, followed by a detailed description of the proposed novel algorithm for the optimal design of automated machine learning models (including NAR, WNNs, and LSTMs).

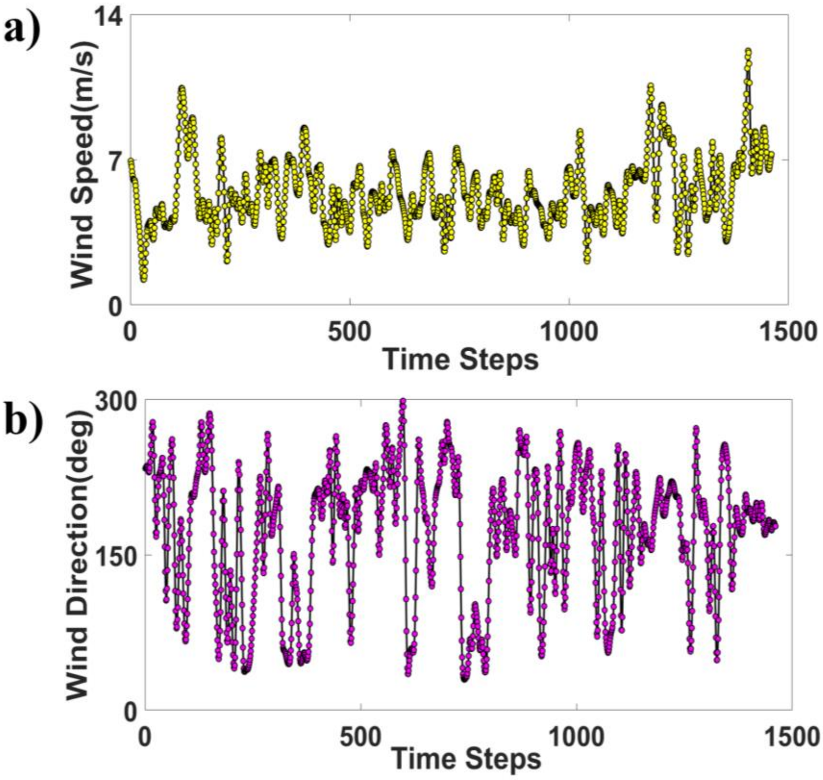


Figure 3. Pictorial representation of 1-year wind characteristics data.

Section 3 describes the results of the proposed work followed by Section 4, which summarizes the conclusions of this work and presents the future scope.

2. Formulation

2.1 Data description and analysis

Wind characteristics data were collected from a French electricity utility company called ENGIE (*La Haute Borne Data| ENGIE*, 2020) over four years with a 6-hour resolution. This data was measured from four wind turbines placed in the corners of a rectangular wind farm of 9 sq. Km area in La Haute Borne, France. Due to the placement of turbines only in the corners, resulting in a large inter-turbine distance, it is assumed that the measured data is not affected by wake effects. Figure 3 gives a pictorial representation of collected wind time-series data on a limited timeframe.

2.1.1. Time-series analysis

Let the data at each time step t be denoted by $\mathrm{X}^{\mathrm{t}} = [\mathrm{X}_1^{\mathrm{t}}\ \mathrm{X}_2^{\mathrm{t}} \ldots \mathrm{X}_{\mathrm{K}}^{\mathrm{t}}]\ \forall\ \mathrm{t} = 1{:}\,\mathrm{T}$, where K is the dimensions and T is the length of the data. In this work, the available data is modeled as two univariate time series, corresponding to wind speed and direction, respectively. Thus, in the current work, K = 1. A nonlinear relationship between input and output variables along with the irregular temporal behavior makes the nature of time-series data nonlinear, which can be detected using a hypothesis test, called the Brock-Dechert-Scheinkman (BDS) test (Akintunde et al., 2015). If the statistical properties such as mean and variance of the time series do not depend on time, then the time series is said to be stationary. A hypothesis test, called Augmented-Dickey-Fuller (ADF) test, is used to determine stationarity in the data (Dickey & Fuller, 1979). The long-term dependency test determines the extent of dependency of data at time instance t on the previous values. The presence of long-term dependency can be determined using the Hurst exponent analysis (Kalo et al., 2019). The results of time series analysis on considered wind data are presented in Section 3.

2.1.2. Time-series Decomposition

The time-series data consists of hidden patterns that have a sequential influence on the data points. A common way of determination proposed by several authors in literature is to decompose the time-series data into the trend, cycle, and seasonal patterns (Guignard et al., 2019). There are several ways to decompose time-series, such as classical-additive, multiplicative, X11, and STL (Hyndman & Athanasopoulos, 2018). Recent literature survey revealed that among others, STL decomposition is most efficient and robust (Hyndman & Athanasopoulos, 2018). Therefore, we consider the STL method for decomposing wind time-series data. The decomposition in the STL method is done through two loops (Cleveland et al., 1990). The outer loop assigns the robustness weights to each data point for Loess smoothing, while the inner loop performs the decomposition. The results are presented in Section 3.

2.2 Methods for modeling nonlinear time-series data.

In univariate time-series modeling, an estimate of data at time step t, $\widehat{\mathrm{X}}^{\mathrm{t}}$, is calculated as a function of B^{T} previous data points: $\mathrm{X}^{\mathrm{p}}|_{\mathrm{p\,=\,t-B^T\ to\ t-1}}$ and a set of tunable parameters θ, which are optimized to minimize the error/loss (L) between the original variable, X^{t} and estimate, $\widehat{\mathrm{X}}^{\mathrm{t}}$, measured $\forall$ t up to the sequence length T. This exercise is called training the time-series model and it is illustrated in Eq. (1) to Eq. (3).

$$\widehat{\mathrm{X}}^{\mathrm{t}} = \mathrm{f}\left(\mathrm{X}^{\mathrm{p}}|_{\mathrm{p\,=\,t-B^T\ to\ t-1}}\ \mathrm{and}\ \theta\right) \tag{1}$$

$$\mathrm{L} = \frac{1}{\mathrm{T-B^T}}\sum_{\mathrm{p=t}}^{\mathrm{T}}\left(\mathrm{X}^{\mathrm{p}} - \widehat{\mathrm{X}}^{\mathrm{p}}\right)^2 \tag{2}$$

$$\theta^* = \mathrm{argmin}(\mathrm{L}) \tag{3}$$

In Eq. (1), f is a functional map. In this manuscript, three nonlinear functional maps are utilized for modeling the wind time-series data. In the first case, f is represented by a neural network regressing on previous data points, thus called a nonlinear autoregressive(NAR) model (Boussaada et al., 2018; Diaconescu, 2008). The θ is the set of weights and biases in the neural network. The description about NAR model is described briefly in Appendix A. In the second case, f is a wavelet neural network(WNN) and θ is the set of translational and dilational parameters (Alexandridis & Zapranis, 2013). The description about WNN model is described briefly in Appendix B. In the third case, f is a deep recurrent neural network called LSTM network and θ is the set of weights and biases in the LSTM network. The description about LSTM is described briefly in Appendix C. In this work, we compare and contrast each of these methods for their abilities and disabilities to model nonlinear time-series data such as wind speed and direction. We also discuss the problems associated with each of these models and present a novel algorithm to alleviate them optimally. The idea behind the proposed algorithm is described below.

2.2.1. Hyper-parameters and idea behind the novel algorithm

In this study, multi-layered (stacked) networks are used in the case of NAR, WNN, and LSTM models for modeling the wind characteristics data. Therefore, before these models are trained, certain hyper-parameters which govern these models need to be fixed. These hyper-parameters include the number of hidden layers in the network, the number of nodes in each hidden layer, activation function, B^T, and learning rate. Conventionally, these hyper-parameters are fixed heuristically allowing severe inaccuracies and limiting the potential of these models. while modeling the NAR, WNN, and LSTM networks. In this work, while the ADAM algorithm (Kingma & Ba, 2014) ensures proper tuning of the learning rate, all other hyper-parameters are estimated optimally using a novel evolutionary multi-objective algorithm. The proposed algorithm is based on bias versus variance trade-off in machine learning:– a simpler model with less number of parameters will have more bias for usage due to its simplicity and high variance in error due to its incapability.

2.3 Algorithm for the optimal design of networks

Utilizing the trade-off between the complexity of the model (in terms of the number of parameters and order of the model B^T) and prediction accuracy (in terms of R^2 on the test set), we present a multi-objective optimization formulation with the objectives of minimizing the complexity of the network and maximizing the accuracy of the model simultaneously. All the hyper-parameters: number of hidden layers, nodes, activation choice and B^T, serve as decision variables. Since the objectives are nonlinear and decision variables are integral, the proposed framework becomes a multi-objective Integer Nonlinear Programming (INLP) problem, which we solved using binary-coded Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb, 2001). The INLP formulation is presented in Eq. (4) and the algorithm is presented in Table 1.

$$\underset{\{N^m:\, m=1:M_{UB}\},\;\; B^T \text{ and } A}{\text{minimize}} -R^2, N_P \text{ and } B^T \tag{4}$$

where,

$$R^2 = \left(\frac{\text{covariance(original and predicted data)}}{\sqrt{\text{var(original data) var(predicted data)}}}\right)^2$$

$$\text{covariance} = \bar{T}\sum_{t=1}^{\bar{T}}\left(X^t\hat{X}^t\right) - \sum_{t=1}^{\bar{T}}\left(X^t\right)\sum_{t=1}^{\bar{T}}\left(\hat{X}^t\right)$$

$$\text{variance} = \overline{T}\sum_{t=1}^{\overline{T}}(X^t)^2 - \left(\sum_{t=1}^{\overline{T}}(X^t)\right)^2$$

N_P is the number of parameters in the network

such that,

$$B^T_{LB} \leq B^T \leq B^T_{UB} \text{ and } N_{LB} \leq N^m \leq N_{UB} \text{ where } N_{LB} = \begin{cases} 1, \text{if } m = 1 \\ 0, \text{if } m > 1 \end{cases}$$

$$A \in \{1,2\} \mid \text{if } A = \begin{cases} 1, \text{ tansigmoid for NAR and LSTMs or Mexican Hat for WNNs} \\ 2, \text{logsigmoid for NAR and LSTMs or Morlet for WNNs} \end{cases}$$

$\{B^T_{LB}, B^T_{UB}, N_{UB}, M_{UB}\} \in \mathbb{Z}_+$ and

M: Number of hidden layers (determined in the proposed algorithm)

M_{LB}, M_{UB}: Lower and upper bound on M (values to be defined a priori – see Table 2)

Table 1. Algorithm for optimal design of NAR, WNN and LSTM models

Step 1	Initialize the number of binary variables as $\mathbf{M_{UB}}$+2 and real variables as 0.
Step 2	Set the parameters of NSGA-II and start the algorithm.

Step 3	For a given population, initialize M = 0 and use the first M_{UB} variables to build the architecture and the last 2 decision variables to determine A and B^T:
	for m → 1 to M_{UB}
	if (decision variable m is not 0) then
	set N^m = decision variable m
	M = M + 1.
	else
	exit the loop
	end if
	end for loop.

Step 4	Assign architecture as [1, {N^m: m = 1 to M}, 1] and obtain B^T and A from remaining two decision variables.
Step 5	Train and validate the model using backpropagation or t-BPTT and ADAM.
Step 6	Test the model using a test size of $\overline{T}$ and evaluate R^2.
Step 7	Evaluate the total parameters N_P in the given network.
Step 8	Increment the population counter, go to step 2 and repeat till a generation is evaluated.
Step 9	Perform the operations of NSGA-II: Crossover, Mutation, Selection and sorting and create the new generation.
Step 10	Repeat Step 3 to 9 till convergence of NSGA-II.

N^m: Number of nodes in hidden layer m (determined in the proposed algorithm)

N_{LB}, N_{UB}: Lower and upper bounds on N^m (values to be defined a priori – see Table 2)

B^T: Length of the unrolled network in case of LSTMs (see Supplymentary file) or the order of NAR and WNN models. (determined in the proposed algorithm and objective)

B^T_{LB}, B^T_{UB}: Lower and upper bounds on B^T (values to be defined a priori – see Table 2)

A: Choice of the activation function (determined in the proposed algorithm)

$\overline{T}$: Number of test data points (values to be defined a priori – see Table 2)

All the models and the optimizer NSGA-II have been coded in Fortran 90 language without the use of any open-source libraries. The simulations are run on Intel® Xeon CPU E5-26900 @ 2.90GHz dual processor 128 GB RAM workstation.

3. Results and Discussions

As the wind time-series data was collected from anemometers, it was first processed using a 5-point moving average approach to remove the measurement noise present in the data. The autocorrelation plot of the residual is shown in Figure 4. The presence of data only between the 95% confidence lines indicates that the residual is white noise (Hyndman &

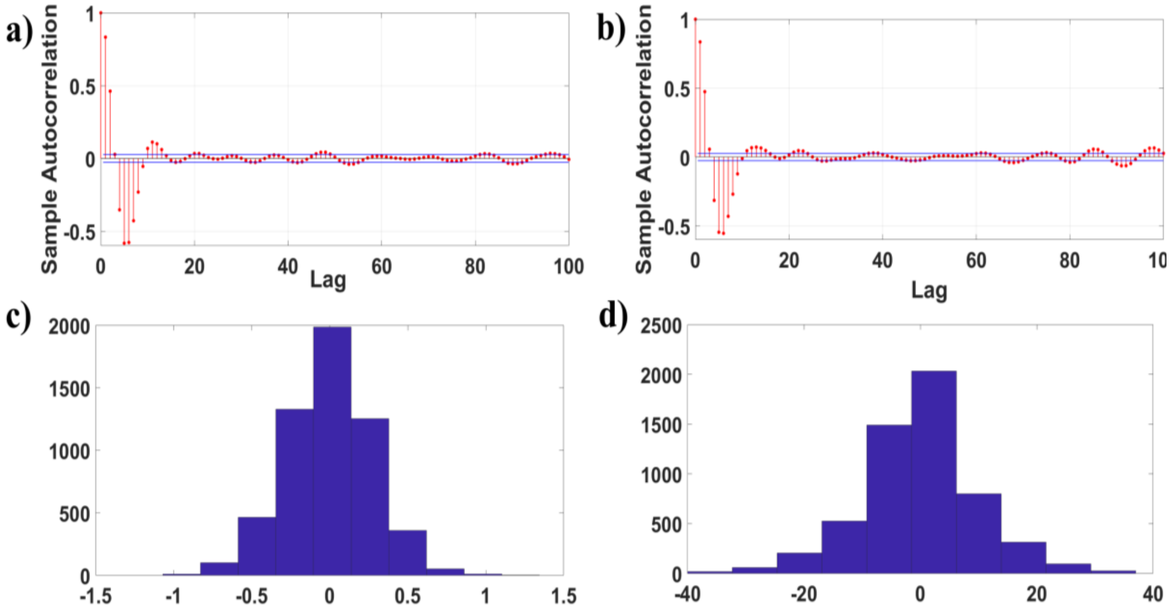


Figure 4. Subfigures (a) and (b) represent the Autocorrelation plots and Subfigures (c) and (d) represent the histograms for residuals in wind speed and direction, respectively.

Athanasopoulos, 2018). To confirm that the data used for plotting the subfigures 4a and 4b is white noise, we also plotted the histograms for them as shown in subfigures 4c and 4d. These Gaussian histograms confirm that the data is indeed white noise. We now present the results of time-series analysis followed by the decomposition of wind data and optimal design of NAR, WNN, and LSTM models.

3.1 Hypothesis tests for Time-series analysis

The nature of time-series data expressed as nonlinearity, stationarity, and long-term dependencies, was examined to determine the appropriate technique for modeling the data. The characteristic of nonlinearity in the data was examined by the BDS test. The null hypothesis is H_0: The time-series data is linear, while the alternate hypothesis is H_1: The time-series data is nonlinear. The level of significance was taken as 5%. The p-values for the wind speed and direction data using the BDS test were observed as 0.001 and 0.003, respectively, which were less than the level of significance. Therefore, the null hypothesis was rejected in both cases and the considered wind time-series data is statistically inferred to be nonlinear.

Similarly, the characteristic of stationarity was examined using the ADF test, where the null hypothesis is H_0: The wind time-series data has Unit root, while the alternate hypothesis is H_1: The time-series data is stationary. The p-values for wind speed and wind direction were reported as 0.008 and 0.01 respectively. Hence, the null hypothesis is rejected and the considered data is inferred statistically to be stationary.

The long-term dependencies in the data were examined using the Hurst Exponent analysis. The Hurst exponent H was determined using the rescale range analysis on periods of observed data. The H values for wind speed and wind direction were observed as 0.83 and 0.79 respectively. As the H values are in the range of 0.5 and 1, the considered data of wind speed and direction is inferred to contain long-term dependencies.

3.2 Time-series decomposition using STL method

The time-series data has hidden patterns, which influence the sequence in the data. Therefore, it was decomposed into three components: trend, seasonal, and remainder/residual, to determine different behavioral patterns using STL decomposition. The results of STL decomposition for wind speed and direction are shown in Figure 5 and Figure 6, respectively. After decomposition, trend and remainder components are modeled together using NAR, WNN, and LSTM networks. Later, seasonality is combined with the forecasts as per STL decomposition for further analysis.

3.3 Optimal design of NAR, WNN, and LSTM models

Once the time-series data of four years is decomposed to extract the seasonality, it is divided into 3-years and 1-year data. The modeling (training and validation) is performed with 3-years data. The trained models are then used to forecast data for the next 1-year, which is compared with the left-out original 1–year data to prove the credibility of forecasts. From the 3-years data, 70% is used for training the models and the remaining 30% is used for validating the trained models. For each of the three models (NAR, WNN, and LSTMs), two parallel simulations of the proposed evolutionary NAS algorithm are run for modeling wind characteristics; one for speed and the other for direction. All the values of bounds on the decision variables (see Eq. (4)) and settings of NSGA-II are listed in Table 2.

Three dimensional Pareto Fronts as solutions were obtained within the first 6-8 generations of the NSGA-II. To confirm the convergence, NSGA-II was also run with different initializations and far more generations than that listed in Table 2. The obtained solutions are presented in Figures 7 and 8 for wind speed and direction, respectively. It can be seen that there lies a trade-off between model accuracy and overfitting as moving from one point to the other in the Pareto front improves one objective at the cost of the other. Decision variables corresponding to these points are called Pareto solutions.

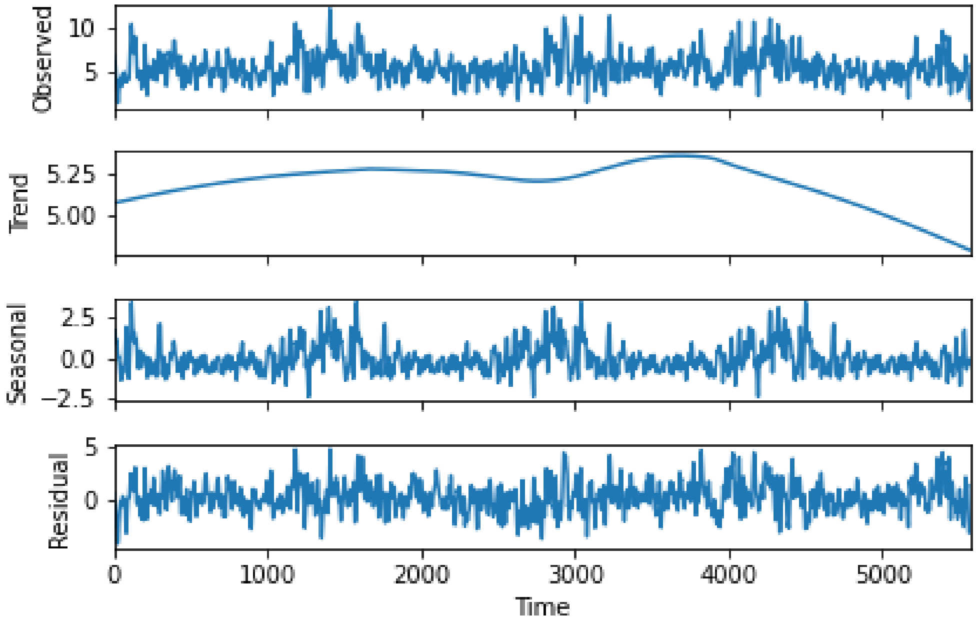


Figure 5. STL Decomposition for Wind Speed

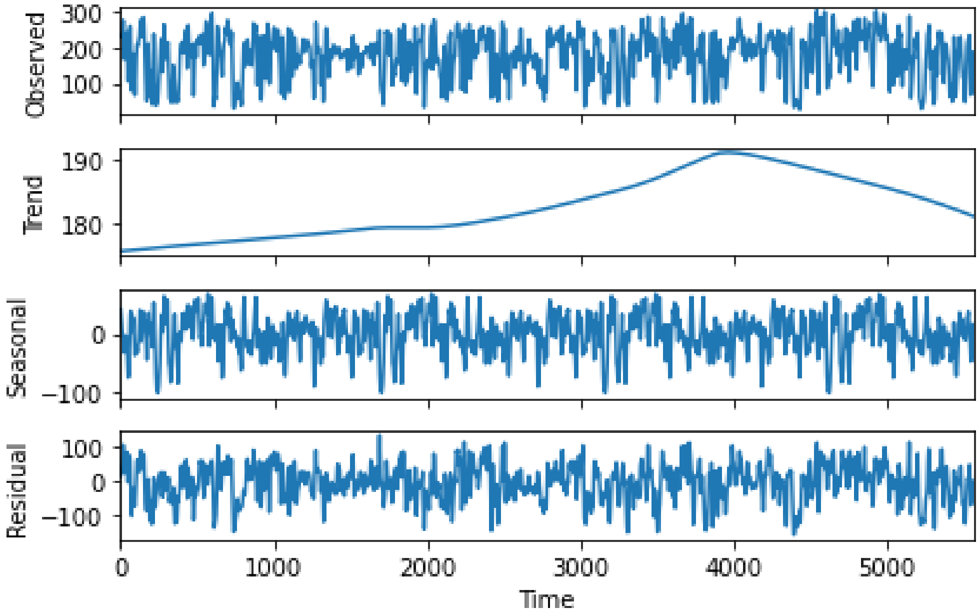


Figure 6. STL Decomposition for Wind Direction

Table 2. List of parameters used in proposed algorithm for evolutionary NAS

S. No	Parameter	Value
1	Number of binary and real variables in NSGA-II	5 and 0
2	Number of population and generations in NSGA-II	200 and 100
3	Mutation and Crossover Probability in NSGA-II	0.01 and 0.9
4	M_{LB}, M_{UB}: Lower and upper bound on number of hidden layers	1 and 3
5	N_{LB}, N_{UB}: Lower and upper bounds on nodes in each hidden layer	{1,0,0} and {16, 15, 15}
6	B^T_{LB}, B^T_{UB}: Lower and upper bounds on B^T	2 and 65
7	$\overline{T}$: Number of test data points	1200

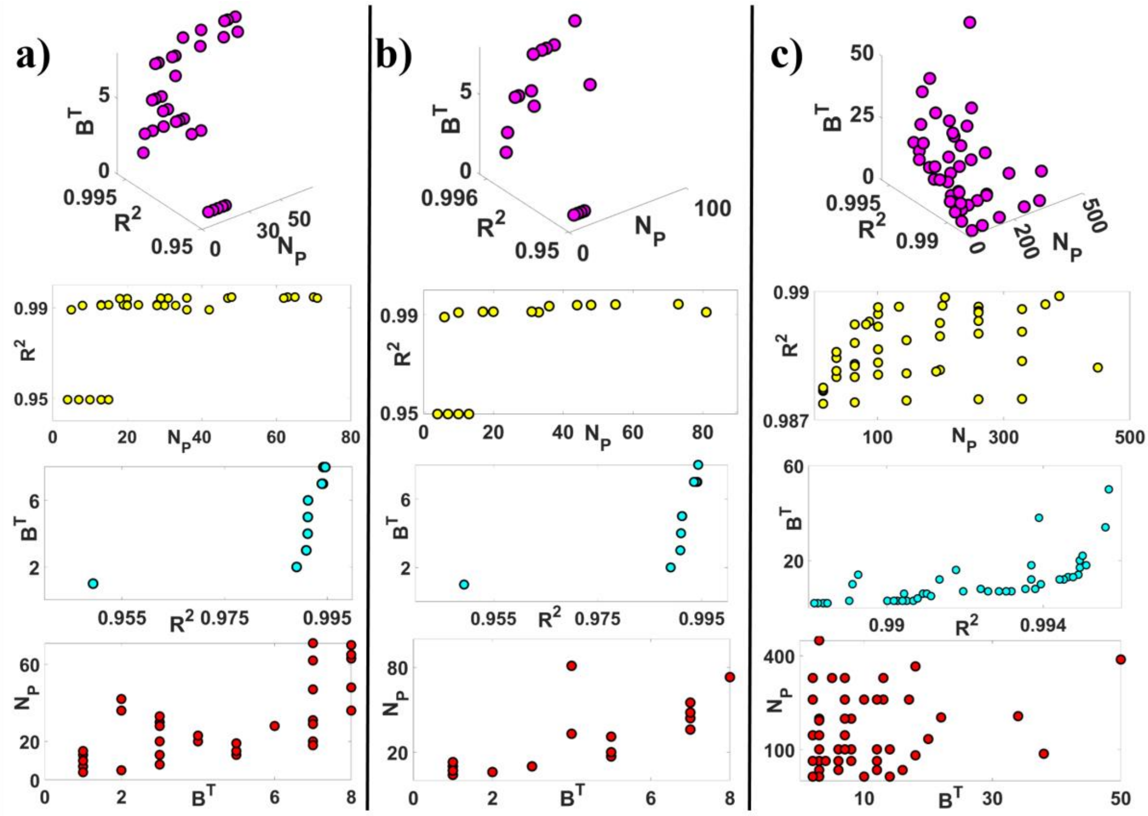


Figure 7. Converged Pareto fronts for (a) NAR, (b) WNN and (c) LSTM for Wind Speed. Row 1 shows the 3D Pareto front while Rows 2 to 4 show the 2D projections.

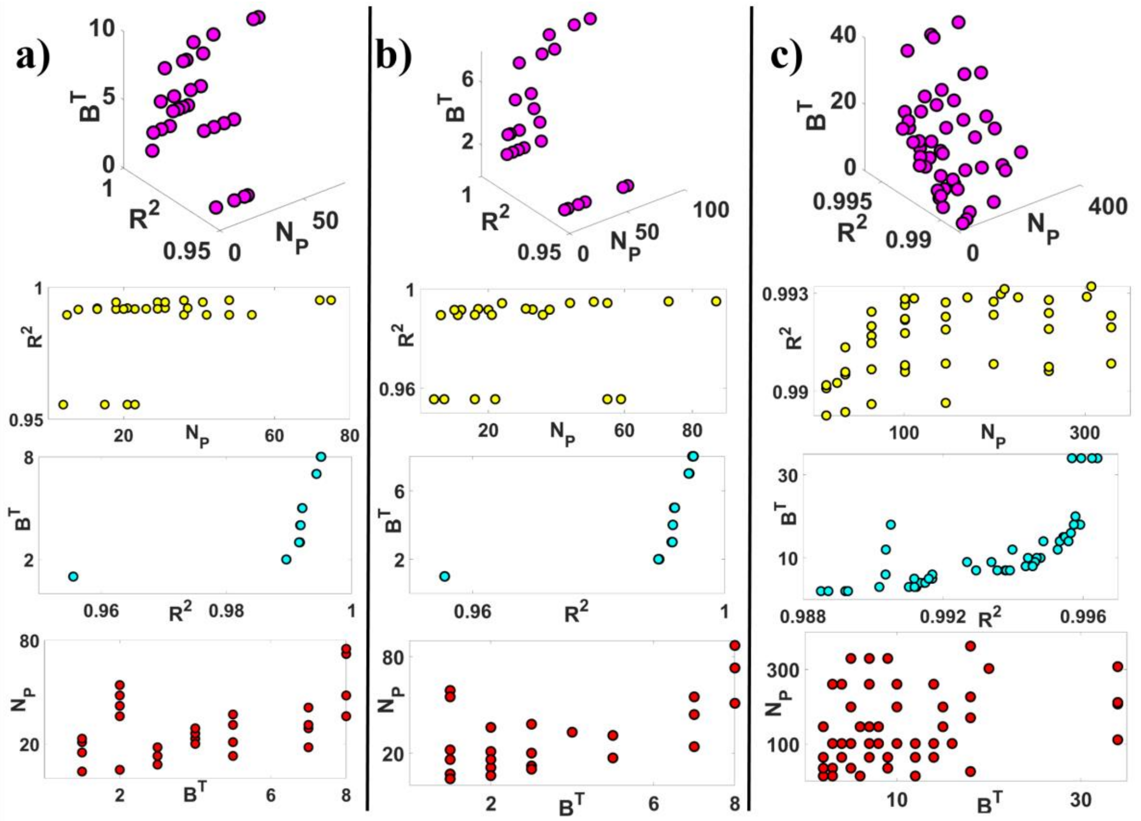


Figure 8. Converged Pareto fronts for (a) NAR, (b) WNN and (c) LSTM for Wind Direction. Row 1 shows the 3D Pareto front while Rows 2 to 4 show the 2D projections.

Each of these solutions is an embodiment of a distinct architecture of NAR, WNN, and LSTMs. The Pareto solutions are shown in Tables D.1 to D.6 in the Appendix D. A single solution from the Pareto list is selected using the Akaike Information Criterion (Akaike, 1987), a robust model selection method, as shown in Eq. (5), which penalizes the models for an increase in the number of parameters thus filtering the overfitted models. Among all models, the one with the least AIC value is selected (Akaike, 1987). The utilization of AIC that ensures the selection of models with less complexity for deciding the final candidate from the list of Pareto solutions once again reinforces the applicability of Green Deep Learning (Xu et al., 2021) in the proposed algorithm. The optimal NAR, WNN, and LSTM models obtained for emulating wind speed and direction, are shown in Table 3. Figures 9 and 10 present the performance of these models for emulating wind speed and direction, respectively.

$$\mathrm{AIC} = \text{Sample size for training} * \log(\mathrm{RMSE}^2) + 2 * \text{Number of parameters} \qquad (5)$$

3.4 Comparisons and Discussions

- As can be seen in Table 3, two and three hidden layered architectures have emerged as the best solutions. Finding them through heuristics would have been extremely time-consuming and laborious. For instance, the proposed algorithm was able to determine the

Table 3: Optimal NAR, WNN and LSTM models obtained from the list of Pareto solutions by implementing the AIC criteria (See Tables D.1 to D.6 in Appendix D for AIC values of each solution)

Speed					
Model	Architecture	Activation Function	Unrolling Length	RMSE on Validation set	R^2 on Validation set
NAR	[1-2-3-4-1]	2	8	0.0102	0.9946
WNN	[1-2-3-4-1]	1	8	0.0105	0.9943
LSTM	[1-5-2-1]	1	34	0.0677	0.9956
Direction					
Model	Architecture	Activation Function	Unrolling Length	RMSE on Validation set	R^2 on Validation set
NAR	[1-2-3-4-1]	2	8	0.0118	0.9951
WNN	[1-2-3-4-1]	1	8	0.0119	0.9950
LSTM	[1-1-6-1]	1	34	0.0623	0.9963

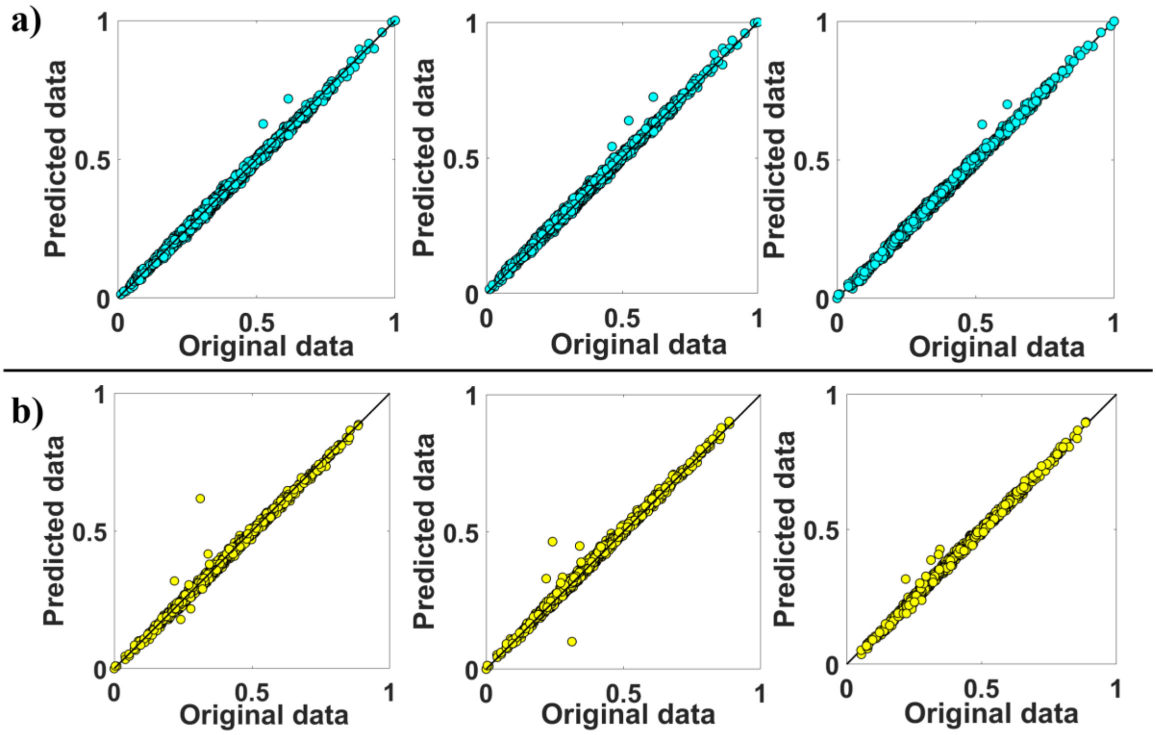


Figure 9. Parity plots of wind direction (a) training data and b) validation data for NAR (column 1), WNN (column 2) and LSTMs (column 3).

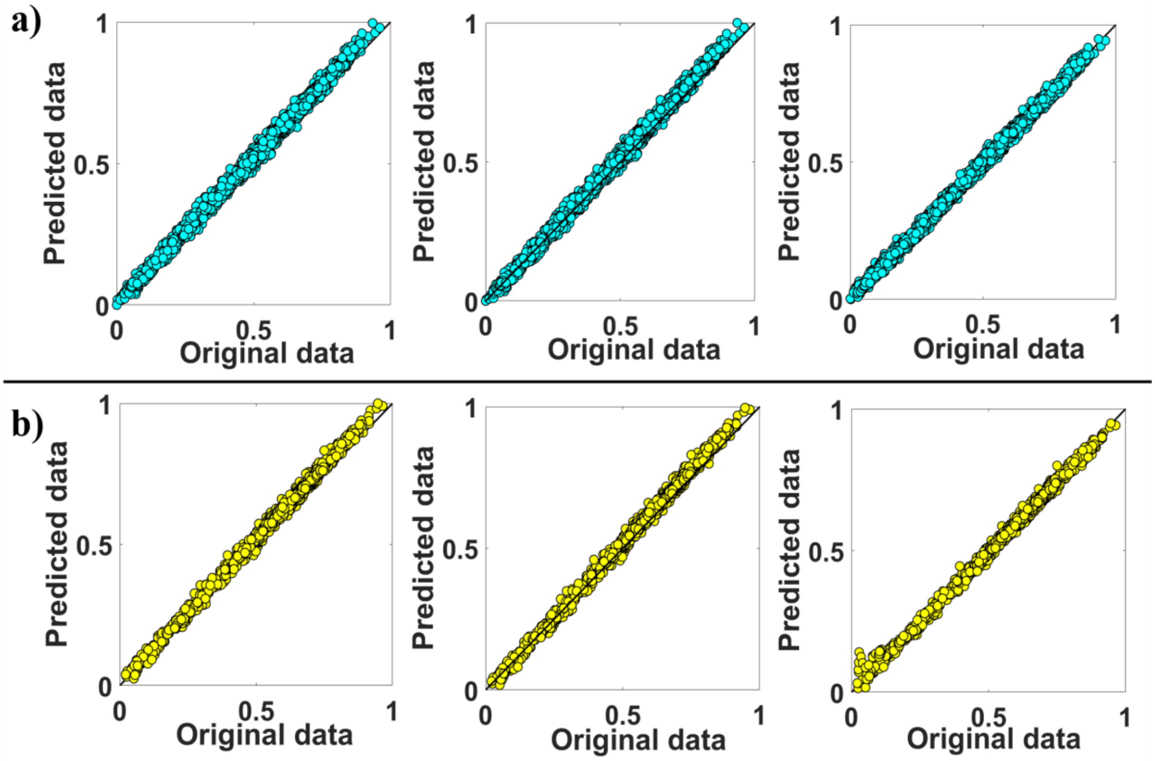


Figure 10. Parity plots of wind direction (a) training data and b) validation data for NAR (column 1), WNN (column 2) and LSTMs (column 3).

best solution from a total of 524288 alternatives (16^3x64x2) by only evaluating a maximum of 521 architectures(maximum obtained in case of LSTM among all three varieties for wind speed. Emulating this with heuristics would have been an impossible task.

- It can be inferred from results in Table 3 that LSTMs have more parameters than NAR and WNN models, however, the accuracy in predicting the training and validation data remains similar. Further, the number of previous time steps required for modeling wind characteristics data is also higher for LSTMs when compared with the other two models. This speaks about the superiority of NAR and WNN models while emulating the training and validation data (see Figures 9 and 10).
- LSTM networks are known for capturing long-term dependencies in the data. And for this functionality, a single LSTM block hosts 4 nodes, each of which is similar to 1 RNN node

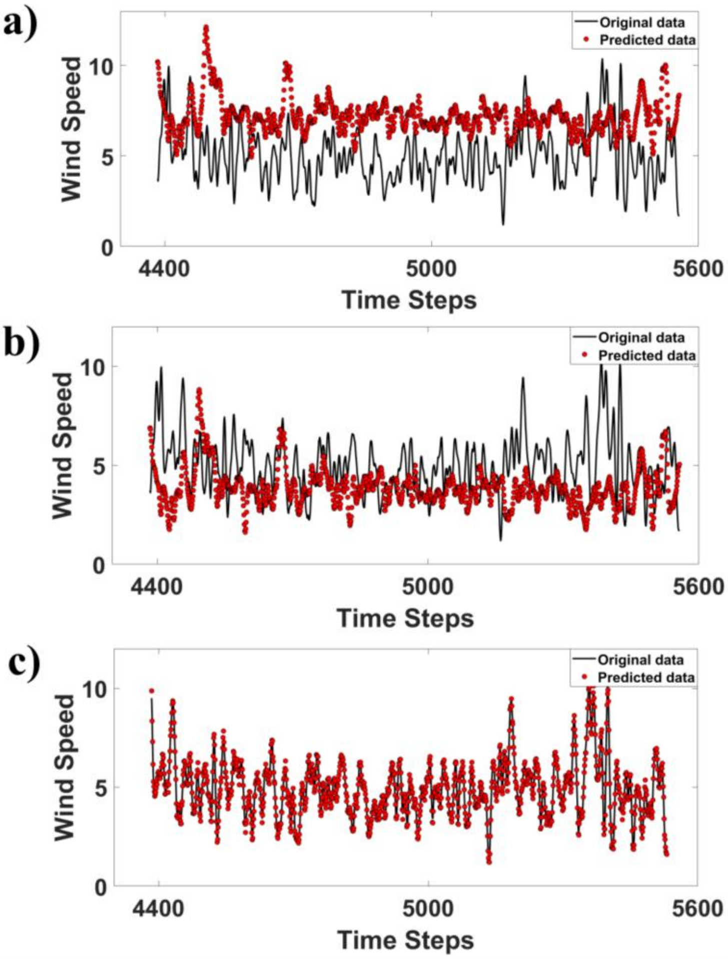


Figure 11. Forecasting of Wind Speed using (a) NAR, (b) WNN and (c) LSTMs

(see Appendix C). This essentially leads to an increase in parameters of the LSTMs as seen in results. At the same time, through Hurst exponent analysis, it was established that the data contains long-term dependencies. Thus, even though, NAR and WNN prove to be simpler and accurate, the capability of LSTMs cannot be undermined. This fact is proven when the three models are compared in terms of forecasts (see Figures 11 and 12).

- The failure of NAR and WNN models to forecast accurately for long-range indicates their inability to capture all dynamic features of the time-series data. On the other hand, accurate forecasts of LSTM justify the necessity of the large number of parameters in them and higher values of B^T. These inferences are also justified from the parity plots in Figure 13. In all three models, B^T is optimally determined and fixed as constant while training. However, only in LSTMs, the extent of dependency on previous data varies to accommodate the dynamics in the data. This is made possible by the forget and input gates which regulate the extent of dependency. Unlike the LSTMs, NAR and WNN models consider the dataset as samples and do not share the parameter information across the timestamps. Hence, these models fail to learn the long-term dynamical behavior in the time-series data. These reasons might have led to the failure of NAR and WNN for forecasting the time-series data over a longer range compared to LSTMs.
- However, when it comes to forecasts over the short range, NAR, WNN, and LSTM models perform similarly. Thus, for applications that demand limited forecasts, optimal NAR and WNN models should be considered rather than LSTM models. This is due to the computational load associated with LSTMs when compared with NAR and WNN models. The importance of short-range forecasts is well established in the domain of wind energy conversion systems (Boussaada et al., 2018). Therefore, we now present a unique application in the design of wind farms, which necessitates long-range forecasts and justifies the applicability of optimal LSTMs.

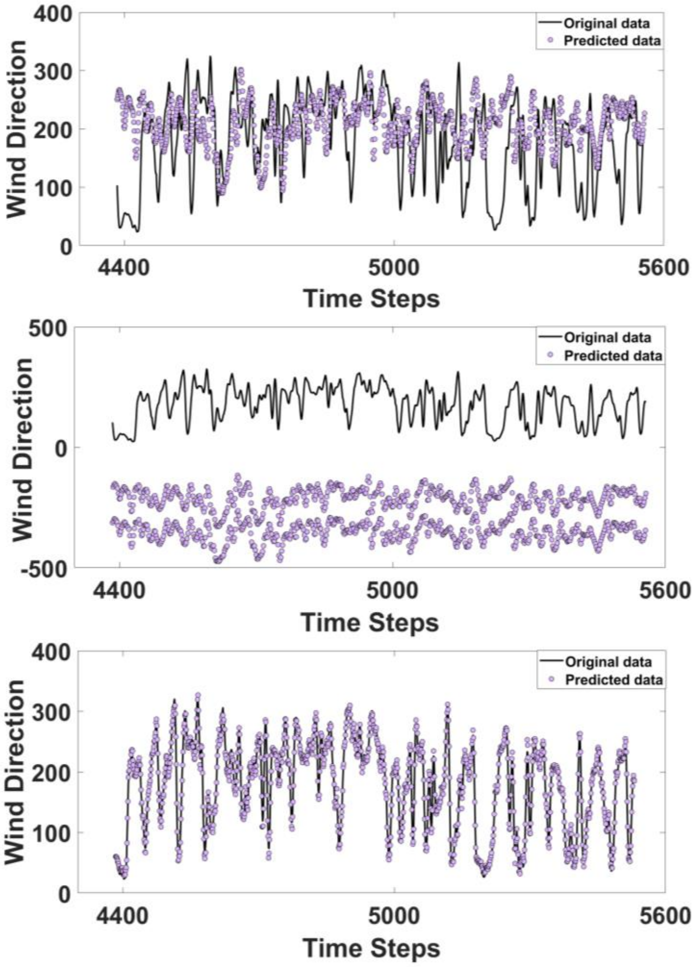


Figure 12. Forecasting of Wind Direction using (a) NAR, (b) WNN and (c) LSTMs.

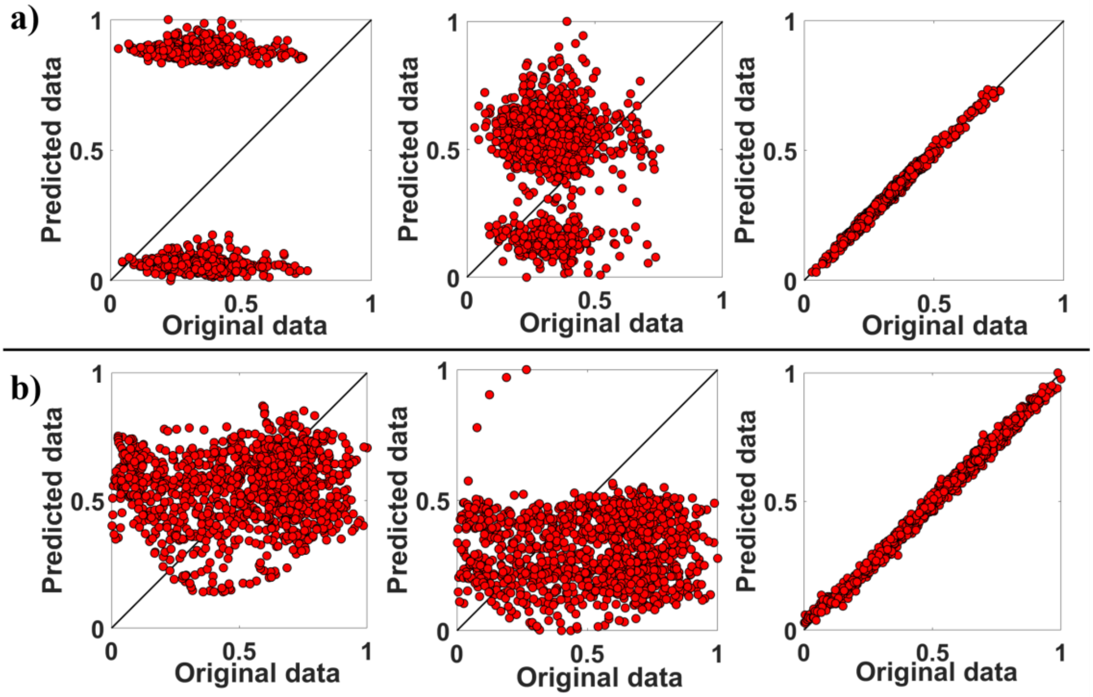


Figure 13. Parity plots of test/forecast data (a) wind speed and (b) wind direction for NAR (column 1), WNN (column 2) and LSTMs (column 3).

3.5 Significance of LSTM forecasts and analysis

Conventionally, energy is extracted from wind using the establishment of a wind farm. An optimal wind farm is where the turbines are arranged in a systematic manner such that the capital expenditure of establishment is minimized and energy obtained from the farm is maximized while considering the wake effects. In this process, called micro-siting, to estimate the energy from a plausible layout,

a) first, a long-range wind time-series data is collected and utilized to construct a Probability Mass Function (PMF) called Wind Frequency Map (WFM),
b) then a suitable method for modeling the wake arising due to the arrangement of turbines is considered to obtain effective velocities at each turbine,
c) the power from each turbine is then evaluated as a function of the effective velocities using a relationship provided by the turbine manufacturer, called the Power curve, and
d) finally, the annual energy from the layout is obtained as a function of the expected value of power from the layout evaluated over the considered WFM as shown in Eq. (6).

$$\mathrm{Energy} = 8760 \sum_{p=1}^{\mathbb{N}} \sum_{q=1}^{D} \sum_{r=1}^{U} \left[P_{curve}\left(u_{effective}(\phi_q, u_r, p) \right) \mathrm{x}\ \mathrm{WFM}(\phi_q, u_r) \right] \quad (6)$$

In Eq. 6, ℕ is the total number of turbines, D is the number of direction sectors, U is the number of speed bins, ϕ_q and u_r are the values of direction and speed in $q^{th}$and r^{th} intervals, respectively, $u_{effective}$ is the effective velocity at a given turbine obtained after application of wake model, P_{curve} is the power curve, the relationship provided by the turbine manufacturer to determine the rated power, and 8760 is the total number of hours in a year. Therefore, it is crucial to construct the WFM accurately (using the frequentist's approach) from the wind time-series data. However, owing to various factors such as unavailability of past wind data due to lack of measurement devices and its archival, unavailability of future wind data due to lack of efficient forecasting techniques, wind farm micro-siting is generally performed based on WFMs constructed using wind characteristics data of shorter duration. Such data depicts minimal or fixed wind characteristics making the wind farm design prone to generate unrealistic estimates of the power. Through this analysis, an effort is made to show the benefit of using accurate and more volume of wind data, while determining the energy production from the wind farms. In what follows next, we first present the procedure for construction of WFM from time-series data, then consider an optimal layout of a wind-farm obtained using micro-siting study (Mittal & Mitra, 2018) and evaluate the annual energy produced from this layout using different WFMs obtained by varying the length of time-series used to build them, for comparison. The wind speed and direction are divided into disjoint intervals (direction sectors and speed bins) which are considered as random variables. The PMF on these random variables is then constructed by the process of counting or the frequentist's approach as shown in Eq. (7).

WFM is a set of discrete probabilities $\mathcal{F}_{\mathrm{ij}} = {}^{\mathrm{T_{ij}}}/_{\mathrm{T}}$ (7)

where T is the total number of points in the wind time-series data and

$\mathrm{T_{ij}}$ is number of points in $\mathrm{i^{th}}$ direction sector and $\mathrm{j^{th}}$ speed bin

In this work, we consider four different WFMs obtained in the following manner by assuming that, we currently have access to the first three-year wind time-series data:

a) $\mathrm{WFM_{aggressive}}$ – the map was constructed by using the most recent year's data.

b) $\mathrm{WFM_{consevative}}$ – the map was constructed by using all three previous years' data.

c) $\mathrm{WFM_{realistic}}$ – the map was constructed by using all three previous years' data and 1-year data forecasted using optimal LSTM obtained in this work and,

d) $\mathrm{WFM_{benchmark}}$ – the map was constructed by using all four years' original data. Since we have assumed that we have access to only three years' data, this map is an ideal case and is created only to show the validity of results obtained in this study. This map serves as the benchmark for comparing the other three WFMs.

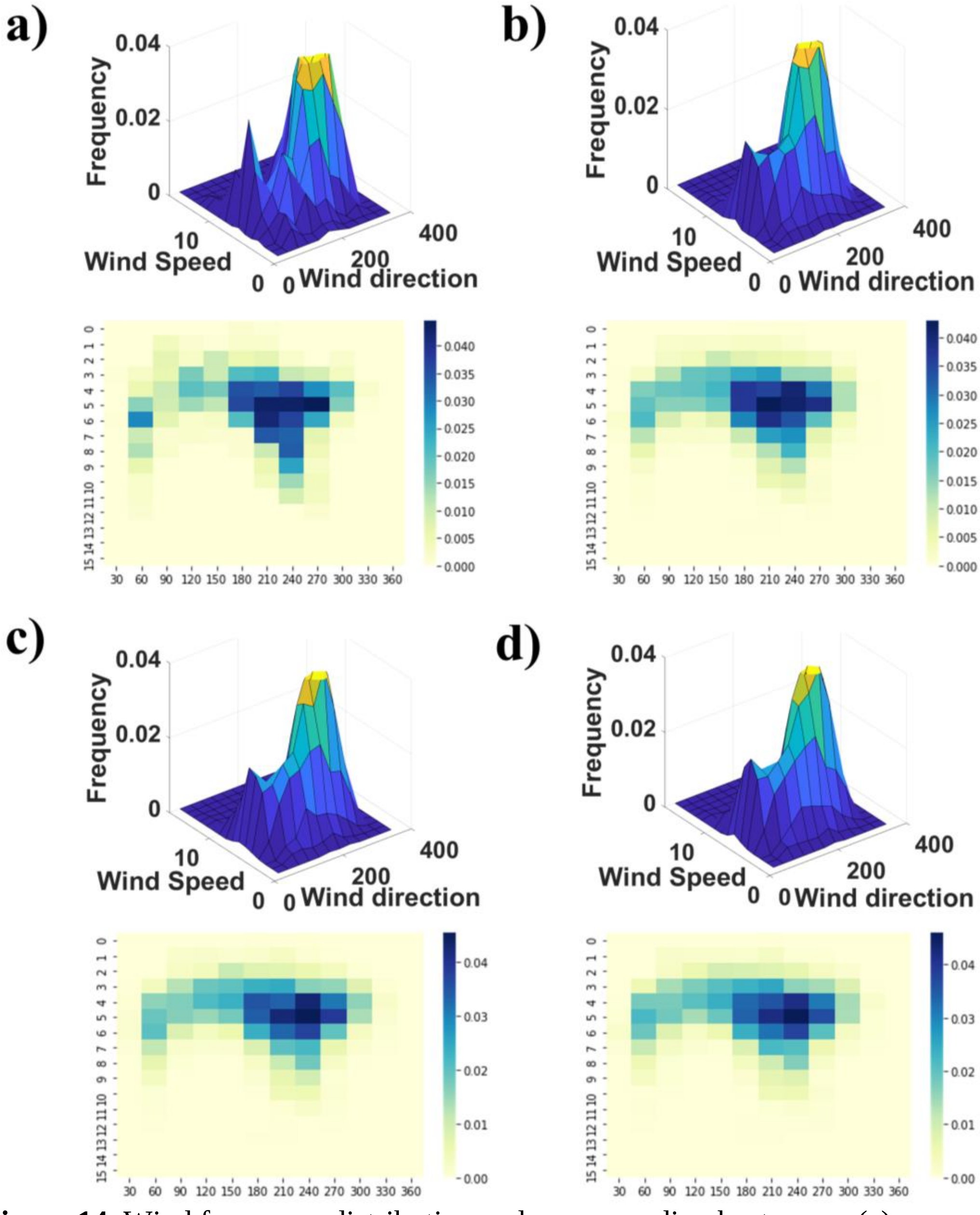


Figure 14. Wind frequency distribution and corresponding heat maps. (a) represents $\mathrm{WFM}_{\mathrm{aggressive}}$. (b) represents $\mathrm{WFM}_{\mathrm{consevative}}$. (c) represents $\mathrm{WFM}_{\mathrm{realistic}}$. (d) represents $\mathrm{WFM}_{\mathrm{benchmark}}$

The frequency maps obtained as described above are shown in Figure 14. To avoid any bias towards the considered WFMs, we use an optimal layout consisting of 33 turbines spread over an area of 3000 sq. Km (see Figure 15), obtained using a micro-siting simulation as described above(Mittal & Mitra, 2018). We then evaluate the expected power from the layout

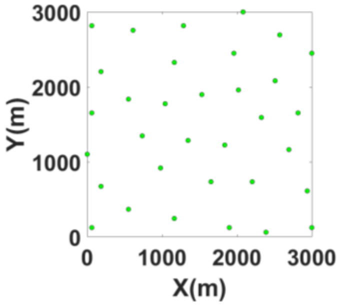


Figure 15. Optimal Wind-farm layout used in this work for analysis using forecasts over long-range of time. The shaded circles indicate turbine locations.

Table 4. Values of Power calculated using the frequency maps obtained from given data and forecasted data.

S. No	Frequency	AEP (kW)
1	$\mathrm{WFM}_{\mathrm{aggressive}}$	14739.92
2	$\mathrm{WFM}_{\mathrm{consevative}}$	12273.48
3	$\mathrm{WFM}_{\mathrm{benchmark}}$	11548.02
4	$\mathrm{WFM}_{\mathrm{realistic}}$	11398.84

as shown in Eq. (6). The values of annual energy obtained for the four different WFMs mentioned previously are listed in Table 4.

The analysis has revealed several interesting insights. Firstly, the energy values for the same layout are widely different with different WFMs. Therefore, if the entire micro-siting study is performed using these layouts, it would reveal entirely different optimal layouts. This would lead to the question of which of them is the correct estimate of the original power. Again the analysis conducted here provides the solution. It shows that the WFM constructed using the combination of original and forecasted data of 4 years has resulted in energy, which is in close approximation to the energy resulted from the benchmark WFM constructed using original data. This justifies the necessity of accurate long-range forecasts of wind characteristics for efficient modeling and simulation of wind energy conversion systems.

4. CONCLUSION

In this work, we compared three state-of-the-art models from the domains of nonlinear system identification and deep learning in terms of their abilities to model and forecast the wind characteristics time-series data. In this process, first the wind characteristics time-series data is analyzed for nonlinearities, non-stationarity, and long-term dependencies and then decomposed to remove the seasonal component from the data. Then, the justification for selecting the NAR, WNN, and LSTM models is presented and the problems associated with their heuristic-based design are articulated. To resolve these issues, a novel evolutionary neural architecture search strategy in the lines of automated machine learning is proposed in this study to optimally design NAR, WNN, and LSTM models. The proposed algorithm not only estimates the hyper-parameters of the models but also ensures the optimal design is driven by the objective to minimize the carbon footprint involved in training and inferring large and deep neural networks. Finally, the significance of accurate forecasts over a long range of time is presented using a study of annual energy production from an optimally designed wind farm. The work is summarized as follows:

- The proposed algorithm provides the best architectures in terms of Pareto solutions which give information about the hyper-parameters of the model. The number of Pareto points from the proposed algorithm was reported as 38, 22, and 45 for NAR, WNN, and LSTM, respectively, for speed. Similarly, the Pareto points for wind direction were reported as 36, 33, and 48 for NAR, WNN, and LSTM, respectively. From the obtained solutions, one Pareto point is selected using the AIC criterion, where the point with minimum AIC is selected for further analysis to prevent overfitted models.
- The training accuracy for modeling wind speed is reported by calculating a statistical metric R^2. The R^2 values for NAR, WNN, and LSTM were reported as 0.9946, 0.9943, 0.9956 respectively, for wind speed. The R^2 values for NAR, WNN, and LSTM were reported as 0.9951, 0.9950, 0.9963 respectively, for wind direction. The results have shown that all three models have performed well while training them.

- However, when compared on the test data over a long range, NAR and WNN models have failed with high RMSE values and very low R^2 values. But LSTM did well on test data for both wind speed and direction with ~99% accuracy.
- Compared to the deep learning models, the system identification techniques do not share the parameters across the timestamps leading to their failure in learning the long-term dependencies in the data. However, for applications requiring predictions for short-range, system identification techniques are more suitable due to their less complexity in terms of model parameters.
- While exploring different designs of the aforementioned models, the proposed algorithm creates a balance between overfitting and parsimony. Though it is shown in this work that the proposed algorithm is capable of designing optimal feedforward and recurrent networks, the idea can also be used to design optimal convolutional networks to model image-based datasets. Thus, the idea is generic and contributes to the novel paradigm of research in machine learning called autoML, aimed at developing automated models without the intervention/implementation of heuristics.
- The significance of accurate forecasting is analysed for improving the annual energy production from an optimally designed wind farm, leading to sustainable clean energy production and a world with near zero carbon foot prints.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by the grant # SPARC/2018-2019/P1084/SL funded by the Ministry of Human Resources Development (MHRD), and grant # DST/NSM/R&D_HPC_Applications/2021/23 funded by the Department of Science and Technology, Government of India for this work.

6. DATA AVAILABILITY

Datasets related to this article can be found at an open-source online repository, ENGIE (*La Haute Borne Data| ENGIE*, 2020).

7. REFERENCES

Abhinav, R., Pindoriya, N. M., Wu, J., & Long, C. (2017). Short-term wind power forecasting using wavelet-based neural network. *Energy Procedia*, *142*, 455–460. https://doi.org/10.1016/J.EGYPRO.2017.12.071

Akaike, H. (1987). Factor Analysis and AIC. In: Selected papers of hirotugu akaike. In *Springer*. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_29

Akintunde, M. O., Oyekunle, J. O., & A., O. G. (2015). Detection of Non-Linearity in the Time Series Using BDS Test. *Science Journal of Applied Mathematics and Statistics*, *3*(4), 184. https://doi.org/10.11648/J.SJAMS.20150304.13

Alexandridis, A. K., & Zapranis, A. D. (2013). Wavelet neural networks: A practical guide. *Neural Networks*, *42*, 1–27. https://doi.org/10.1016/J.NEUNET.2013.01.008

Allen, D. J., Tomlin, A. S., Bale, C. S. E., Skea, A., Vosper, S., & Gallani, M. L. (2017). A boundary layer scaling technique for estimating near-surface wind energy using numerical weather prediction and wind map data. *Applied Energy*, *208*, 1246–1257. https://doi.org/10.1016/J.APENERGY.2017.09.029

An, X., Jiang, D., Liu, C., & Zhao, M. (2011). Wind farm power prediction based on wavelet decomposition and chaotic time series. *Expert Systems with Applications*, *38*(9), 11280–11285. https://doi.org/10.1016/J.ESWA.2011.02.176

Boussaada, Z., Curea, O., Remaci, A., Camblong, H., & Bellaaj, N. M. (2018). A Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation. *Energies 2018, Vol. 11, Page 620*, *11*(3), 620. https://doi.org/10.3390/EN11030620

Brahimi, T. (2019). Using Artificial Intelligence to Predict Wind Speed for Energy Application in Saudi Arabia. *Energies 2019, Vol. 12, Page 4669*, *12*(24), 4669. https://doi.org/10.3390/EN12244669

Chen, J., Liu, H., & Chen, C. (2022). Wind speed forecasting using a novel multi-scale feature adaptive extraction ensemble model with multi-objective error regression correction. *Expert Systems with Applications*, 117358. https://doi.org/10.1016/J.ESWA.2022.117358

Cho, H., Kim, Y., Lee, E., Choi, D., Lee, Y., & Rhee, W. (2020). Basic Enhancement Strategies When Using Bayesian Optimization for Hyperparameter Tuning of Deep Neural Networks. *IEEE Access*, *8*, 52588–52608. https://doi.org/10.1109/ACCESS.2020.2981072

Ciri, U., Santoni, C., Bernardoni, F., Salvetti, M. V., & Leonardi, S. (2019). Development of a surrogate model for wind farm control. *Proceedings of the American Control Conference*, *2019-July*, 2849–2854. https://doi.org/10.23919/ACC.2019.8814766

Cleveland, R., Cleveland, W., McRae, J., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess (with discussion) – ScienceOpen. *Journal of Official Statistics*, *6*, 3–73. https://www.scienceopen.com/document?vid=be074647-46c0-4ba9-991e-3124fbf63ed1

Daniel, L. O., Sigauke, C., Chibaya, C., & Mbuvha, R. (2020). Short-Term Wind Speed Forecasting Using Statistical and Machine Learning Methods. *Algorithms 2020, Vol. 13, Page 132*, *13*(6), 132. https://doi.org/10.3390/A13060132

Deb, K. (2001). Multi-objective optimization using evolutionary algorithms. In *John Wiley & Sons*. John Wiley & Sons. https://www.wiley.com/en-us/Multi+Objective+Optimization+using+Evolutionary+Algorithms-p-9780471873396

Diaconescu, E. (2008). The use of NARX Neural Networks to predict Chaotic Time Series. *WSEAS Transactions on Computer Research*, *3*(3), 182–191.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, *74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Ding, M., Zhou, H., Xie, H., Wu, M., Nakanishi, Y., & Yokoyama, R. (2019). A gated recurrent unit neural networks based wind speed error correction model for short-term wind power forecasting. *Neurocomputing*, *365*, 54–61. https://doi.org/10.1016/J.NEUCOM.2019.07.058

Dong, X., Shen, J., Wang, W., Shao, L., Ling, H., & Porikli, F. (2021). Dynamical hyperparameter optimization via deep reinforcement learning in tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(5), 1515–1529. https://doi.org/10.1109/TPAMI.2019.2956703

Global Wind Report 2021 - Global Wind Energy Council. (2021). https://gwec.net/global-wind-report-2021/

Grubb, M., Koch, M., Thomson, K., Sullivan, F., & Munson, A. (2019). The "Earth Summit" agreements : a guide and assessment : an analysis of the Rio '92 UN Conference on Environment and Development. *Routledge*, *9*.

Guignard, F., Lovallo, M., Laib, M., Golay, J., Kanevski, M., Helbig, N., & Telesca, L. (2019). Investigating the time dynamics of wind speed in complex terrains by using the Fisher–Shannon method. *Physica A: Statistical Mechanics and Its Applications*, *523*, 611–621. https://doi.org/10.1016/J.PHYSA.2019.02.048

Han, J. H., Choi, D. J., Park, S. U., & Hong, S. K. (2020). Hyperparameter Optimization Using a Genetic Algorithm Considering Verification Time in a Convolutional Neural Network. *Journal of Electrical Engineering & Technology 2020 15:2*, *15*(2), 721–726. https://doi.org/10.1007/S42835-020-00343-7

Hyndman, J. R., & Athanasopoulos, G. (2018). Forecasting: principles and practice - Rob J

Hyndman, George Athanasopoulos - Google Books. In *otexts*. https://books.google.co.in/books?hl=en&lr=&id=_bBhDwAAQBAJ&oi=fnd&pg=PA7&dq=%5B42%5D%09Hyndman+RJ,+Athanasopoulos+G.+Forecasting:+principles+and+practice.+OTexts%3B+2018&ots=TiiVtnTQEK&sig=nsM3P4MdPxC_LzM1mhr5zW4XNb4&redir_esc=y#v=onepage&q&f=false

Jahangir, H., Golkar, M. A., Alhameli, F., Mazouz, A., Ahmadian, A., & Elkamel, A. (2020). Short-term wind speed forecasting framework based on stacked denoising auto-encoders with rough ANN. *Sustainable Energy Technologies and Assessments*, *38*, 100601. https://doi.org/10.1016/J.SETA.2019.100601

Kalo, L., Kamalanathan, P., Pant, H. J., Cassanello, M. C., & Upadhyay, R. K. (2019). Mixing and regime transition analysis of liquid-solid conical fluidized bed through RPT technique. *Chemical Engineering Science*, *207*, 702–712. https://doi.org/10.1016/J.CES.2019.07.005

Kingma, D. P., & Ba, J. L. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1412.6980v9

Kumar Dubey, A., Kumar, A., García-Díaz, V., Kumar Sharma, A., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments*, *47*, 101474. https://doi.org/10.1016/J.SETA.2021.101474

[dataset] *La Haute Borne Data| ENGIE*. (2020). https://opendata-renewables.engie.com/explore/

Li, H., Jiang, Z., Shi, Z., Han, Y., Yu, C., & Mi, X. (2022). Wind-speed prediction model based on variational mode decomposition, temporal convolutional network, and sequential triplet loss. *Sustainable Energy Technologies and Assessments*, *52*, 101980.

https://doi.org/10.1016/J.SETA.2022.101980

Li, L. L., Liu, Z. F., Tseng, M. L., Jantarakolica, K., & Lim, M. K. (2021). Using enhanced crow search algorithm optimization-extreme learning machine model to forecast short-term wind power. *Expert Systems with Applications, 184,* 115579. https://doi.org/10.1016/J.ESWA.2021.115579

Liu, X., Lin, Z., & Feng, Z. (2021). Short-term offshore wind speed forecast by seasonal ARIMA - A comparison against GRU and LSTM. *Energy*, *227*, 120492. https://doi.org/10.1016/J.ENERGY.2021.120492

Liu, Z., Jiang, P., Wang, J., & Zhang, L. (2021). Ensemble forecasting system for short-term wind speed forecasting based on optimal sub-model selection and multi-objective version of mayfly optimization algorithm. *Expert Systems with Applications*, *177*, 114974. https://doi.org/10.1016/J.ESWA.2021.114974

Miao, S., Yang, H., & Gu, Y. (2018). A wind vector simulation model and its application to adequacy assessment. *Energy, 148,* 324–340. https://doi.org/10.1016/J.ENERGY.2018.01.109

Mittal, P., & Mitra, K. (2018). Determining layout of a wind farm with optimal number of turbines: A decomposition based approach. *Journal of Cleaner Production*, *202*, 342–359. https://doi.org/10.1016/J.JCLEPRO.2018.08.093

Ningsih, F. R., Djamal, E. C., & Najmurrakhman, A. (2019). Wind Speed Forecasting Using Recurrent Neural Networks and Long Short Term Memory. *Proceedings of the 2019 6th International Conference on Instrumentation, Control, and Automation, ICA 2019*, 137–141. https://doi.org/10.1109/ICA.2019.8916717

Olaofe, Z. O. (2014). A 5-day wind speed & power forecasts using a layer recurrent neural network (LRNN). *Sustainable Energy Technologies and Assessments*, *6*, 1–24. https://doi.org/10.1016/J.SETA.2013.12.001

Prasetyowati, A., Sudibyo, H., & Sudiana, D. (2017). Wind Power Prediction by Using Wavelet Decomposition Mode Based NARX-Neural Network. *ACM International Conference Proceeding Series*, 275–278. https://doi.org/10.1145/3168390.3168434

Renewables in Electricity Production | Statistics Map by Region | Enerdata. (2021). https://yearbook.enerdata.net/renewables/renewable-in-electricity-production-share.html

Salcedo-Sanz, S., Ortiz-García, E. G., Pérez-Bellido, Á. M., Portilla-Figueras, A., & Prieto, L. (2011). Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Systems with Applications*, *38*(4), 4052–4057. https://doi.org/10.1016/J.ESWA.2010.09.067

Song, Z., Geng, X., Kusiak, A., & Xu, C. (2011). Mining Markov chain transition matrix from wind speed time series data. *Expert Systems with Applications*, *38*(8), 10229–10239. https://doi.org/10.1016/J.ESWA.2011.02.063

Trebing, K., & Mehrkanoon, S. (2020). Wind speed prediction using multidimensional convolutional neural networks. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*, 713–720. https://doi.org/10.1109/SSCI47803.2020.9308323

Tso, W. W., Burnak, B., & Pistikopoulos, E. N. (2020). HY-POP: Hyperparameter optimization of machine learning models through parametric programming. *Computers & Chemical Engineering*, *139*, 106902. https://doi.org/10.1016/J.COMPCHEMENG.2020.106902

Vogler, J. (2021). The international politics of COP26. *Scottish Geographical Journal*, *136*(1–4), 31–35. https://doi.org/10.1080/14702541.2020.1863610

Wang, Jianzhou, Li, H., Wang, Y., & Lu, H. (2021). A hesitant fuzzy wind speed forecasting system with novel defuzzification method and multi-objective optimization algorithm. *Expert Systems with Applications*, *168*, 114364. https://doi.org/10.1016/J.ESWA.2020.114364

Wang, Jujie, Gao, D., Zhuang, Z., & Wu, J. (2022). An optimized complementary prediction method based on data feature extraction for wind speed forecasting. *Sustainable Energy Technologies and Assessments*, *52*, 102068. https://doi.org/10.1016/J.SETA.2022.102068

Wang, K., Qi, X., Liu, H., & Song, J. (2018). Deep belief network based k-means cluster approach for short-term wind power forecasting. *Energy*, *165*, 840–852. https://doi.org/10.1016/J.ENERGY.2018.09.118

World Energy Consumption Statistics | Enerdata. (2021). https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html

Wu, J., Chen, S. P., & Liu, X. Y. (2020). Efficient hyperparameter optimization through model-based reinforcement learning. *Neurocomputing*, *409*, 381–393. https://doi.org/10.1016/J.NEUCOM.2020.06.064

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Wu, Y. X., Wu, Q. B., & Zhu, J. Q. (2019). Data-driven wind speed forecasting using deep feature extraction and LSTM. *IET Renewable Power Generation*, *13*(12), 2062–2069. https://doi.org/10.1049/IET-RPG.2018.5917

Xu, J., Zhou, W., Fu, Z., Zhou, H., & Li, L. (2021). *A Survey on Green Deep Learning*. *3*. http://arxiv.org/abs/2111.05193

Xue, H., Jia, Y., Wen, P., & Farkoush, S. G. (2020). Using of improved models of Gaussian Processes in order to Regional wind power forecasting. *Journal of Cleaner Production*, *262*, 121391. https://doi.org/10.1016/J.JCLEPRO.2020.121391

Zhang, S., Wang, C., Liao, P., Xiao, L., & Fu, T. (2022). Wind speed forecasting based on model selection, fuzzy cluster, and multi-objective algorithm and wind energy simulation by Betz's theory. *Expert Systems with Applications*, *193*, 116509. https://doi.org/10.1016/J.ESWA.2022.116509

APPENDIX

Appendix A: Nonlinear Autoregressive Models (NAR):

In case of NAR model, the sequential data is rearranged into input-output data pairs based on the value of $\mathrm{B^T}$. Figure A.1 presents a typical example of this rearrangement with $\mathrm{B^T} = 3$ and corresponding NAR model with neural network as the nonlinear map f. The inputs are processed across the network to generate the estimate of the target (see Eq. (A.1) to Eq. (A.3)), which results in generation of a loss function that can be used to optimize the weights and biases in the network (see Eq. (2) and (3) in the manuscript). To model nonlinearity in the data, the network hosts a set of nonlinear transformation functions called activation functions. Log-sigmoid and Tan-sigmoid functions shown in Eq. (A.4) are two commonly used activation functions in ANNs.

- Evaluation of activated output of ith node in first hidden layer (x_{i}^{1}):

$$y_{\mathrm{i}}^{1} = \sum_{\mathrm{p=t-B^T}}^{\mathrm{t-1}} \left(\mathrm{w_{ij}^{1}\, X^{p}}\right) + \mathrm{b_{i}^{1}}, \text{ where, } \mathrm{j = p - t + B^T + 1} \text{ and } x_{\mathrm{i}}^{1} = \varphi(y_{\mathrm{i}}^{1}) \qquad \text{(A.1)}$$

Here, y is the weighted sum of inputs $\mathrm{X^p}|_{\mathrm{p\,=\,t-B^T\ to\ t-1}}$, w and b are weights and biases, respectively, and φ is the activation function. The superscript (on y, x, w and b) indicates the layer number, the first subscript indicates the current node in the given layer and second

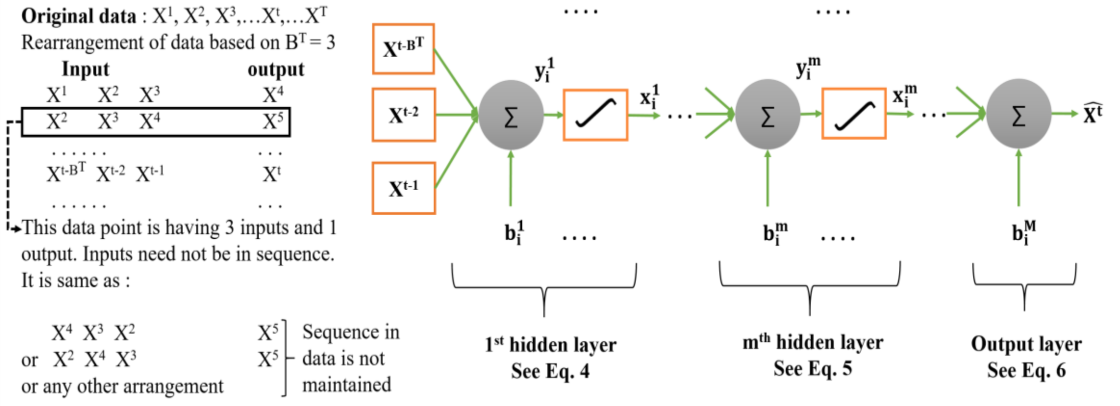


Figure A.1. Pictorial representation of NAR model.

subscript indicates the node in previous layer which is connected to the current node. For example, $\mathrm{w}_{\mathrm{ij}}^{\mathrm{m}}$ indicates the weight on connection from j^{th} node in $(\mathrm{m}-1)^{\mathrm{th}}$ layer to i^{th} node in m^{th} layer.

- Evaluation of activated output of ith node in mth hidden layer (x_i^m):

$$y_{\mathrm{i}}^{\mathrm{m}} = \sum_{\mathrm{j}=1}^{\mathrm{N}^{\mathrm{m}-1}} \left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{m}}\, x_{\mathrm{j}}^{\mathrm{m}-1}\right) + \mathrm{b}_{\mathrm{i}}^{\mathrm{m}} \text{ and } x_{\mathrm{i}}^{\mathrm{m}} = \varphi(y_{\mathrm{i}}^{\mathrm{m}})\ \forall\ \mathrm{m} = 2 \text{ to } \mathrm{M}-1 \qquad \text{(A.2)}$$

Here, $\mathrm{N}^{\mathrm{m}-1}$ is number of nodes in (m-1)th layer and M is total number of layers (hidden layers + output layer) in the network

- Evaluation of network output ($\widehat{\mathrm{X}}^{\mathrm{t}}$):

$$\widehat{\mathrm{X}}^{\mathrm{t}} = \sum_{\mathrm{j}=1}^{\mathrm{N}^{\mathrm{M}-1}} \left(\mathrm{w}_{1\mathrm{j}}^{\mathrm{M}}\, x_{\mathrm{j}}^{\mathrm{M}-1}\right) + \mathrm{b}_{1}^{\mathrm{M}} \qquad \text{(A.3)}$$

- Commonly used activation functions (φ):

$$\text{Log-sigmoid: } \varphi(y) = \frac{1}{1+\exp(-y)} \text{ and Tan-sigmoid: } \varphi(y) = \frac{2}{1+\exp(-2y)} - 1 \qquad \text{(A.4)}$$

Appendix B: Wavelet Neural Networks:

Wavelet neural networks are similar to feedforward neural networks where the sigmoid activation functions are replaced with wavelet functions. In contrast to sigmoid neural networks, the wavelet networks, often considered as generalization of Radial Basis Function (RBF) networks, are efficient in initializing the parameters such that they converge to the global minimum of the error function (Alexandridis & Zapranis, 2013). The given time-series data is rearranged in a similar way as it is done in case of NAR models. In the input layer of wavelet network, the explanatory variables ($\mathrm{X}^{\mathrm{p}}|_{\mathrm{p}\,=\,\mathrm{t}-\mathrm{B}^{\mathrm{T}} \text{ to } \mathrm{t}-1}$) are introduced. Nodes in the hidden

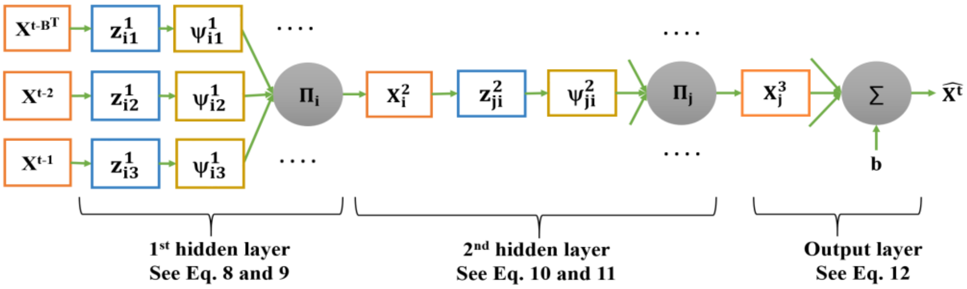


Figure B.1. Pictorial representation of WNN model.

layers, called wavelons, transform the input variables to translated and dilated versions of mother wavelet. The translation controls the position of the mother wavelet and dilation controls the scaling parameter. The output layer approximates the estimated target value. The structure of simple wavelet network with two hidden layers is shown in Figure B.1, and the equations for evaluation of network output are shown in Eq. (B.1) to Eq. (B.5).

- Evaluation of activated output of ith node in first hidden layer (x_{i}^{1}):

$$\mathrm{z}_{\mathrm{ij}}^{1} = \left(\mathrm{X}^{\mathrm{p}} - \mathrm{w1}_{\mathrm{ij}}^{1}\right)/\left(\mathrm{w2}_{\mathrm{ij}}^{1}\right) \text{ and } \Psi_{\mathrm{ij}}^{1} = \varphi\left(\mathrm{z}_{\mathrm{ij}}^{1}\right)$$

$$\forall\ \mathrm{p} = \mathrm{t} - \mathrm{B}^{\mathrm{T}} \text{ to } \mathrm{t} - 1\ \&\ \mathrm{j} = \mathrm{p} - \mathrm{t} + \mathrm{B}^{\mathrm{T}} + 1 \tag{B.1}$$

$$x_{\mathrm{i}}^{1} = \prod_{\mathrm{j}=\mathrm{t}-\mathrm{B}^{\mathrm{T}}}^{\mathrm{t}-1}\left(\Psi_{\mathrm{ij}}^{1}\right) \tag{B.2}$$

Here, z is the input variable translated and dilated using the weights w1 and w2, respectively, Ψ is the output after application of a wavelet transform on z and φ is the wavelet function (see Eq. (B.6)). The superscript (on z, x, w1, w2 and Ψ) indicates the layer number, the first subscript indicates the current node in the given layer and second subscript indicates the node in previous layer which is connected to the current node.

- Evaluation of activated output of ith node in mth hidden layer ($x_{\mathrm{i}}^{\mathrm{m}}$):

$$\mathrm{z}_{\mathrm{ij}}^{\mathrm{m}} = \left(x_{\mathrm{j}}^{\mathrm{m}-1} - \mathrm{w1}_{\mathrm{ij}}^{\mathrm{m}}\right)/\left(\mathrm{w2}_{\mathrm{ij}}^{\mathrm{m}}\right) \text{ and } \Psi_{\mathrm{ij}}^{\mathrm{m}} = \varphi\left(\mathrm{z}_{\mathrm{ij}}^{\mathrm{m}}\right)$$

$$\forall\ \mathrm{j} = 1 \text{ to } \mathrm{N}^{\mathrm{m}-1}\ \&\ \mathrm{m} = 2 \text{ to } \mathrm{M} - 1 \tag{B.3}$$

$$x_{\mathrm{i}}^{\mathrm{m}} = \prod_{\mathrm{j}=1}^{\mathrm{N}^{\mathrm{m}-1}}\left(\Psi_{\mathrm{ij}}^{\mathrm{m}}\right)\ \forall\ \mathrm{m} = 2 \text{ to } \mathrm{M} - 1 \tag{B.4}$$

Here, $\mathrm{N}^{\mathrm{m}-1}$ is number of nodes in (m-1)th layer and M is total number of layers (hidden layers + output layer) in the network

- Evaluation of network output ($\widehat{\mathrm{X}}^{\mathrm{t}}$):

$$\widehat{\mathrm{X}}^{\mathrm{t}} = \sum_{\mathrm{j}=1}^{\mathrm{N}^{\mathrm{M}-1}}\left(\mathrm{w}_{1\mathrm{j}}^{\mathrm{M}}\, x_{\mathrm{j}}^{\mathrm{M}-1}\right) + \mathrm{b}_{1}^{\mathrm{M}} \tag{B.5}$$

- Commonly used wavelet functions (φ):

Mexican hat: $\varphi(\mathrm{y}) = \frac{2}{\sqrt{3}}\pi^{\frac{-1}{4}}(1 - \mathrm{y}^{2})\exp(\frac{-\mathrm{y}^{2}}{2})$ and

$$\text{Morlet: } \varphi(y) = \cos(1.75y)\exp(\frac{-y^2}{2}) \quad \text{(B.6)}$$

Appendix C: Long Short-term Memory Networks

A primary difference between the auto-regressive models, such as NAR and WNNs, and the RNNs is that, while the former regress on previous data points, the recurrent networks regress on previous hidden states as shown in Eq. (C.1) to (C.3). This is made possible using a feedback loop on every node in the hidden layers (see Figure C.1). Several such hidden layers connected in a sequence between input and output layers together constitute the recurrent network. As opposed to a feed forward network, the recurrent network has two dimensions – one goes forward in layers (input to output layer) and the other goes forward in time as shown in Figure C.1. This additional dimension in time is such that the same network is simulated repeatedly with feedback from previous time point and information at current time point. Therefore, to differentiate the network variables from one time-step to another, an additional superscript is added which indicates time (the other superscript indicates layers). However, since the network remains same across all time instances, the parameters, i.e., the weights and biases do not change with time steps. This kind of architecture maintains the sequence in the data while training the model, unlike NAR and WNN models (Alexandridis & Zapranis, 2013). Eq. (C.1) and (C.2) are valid $\forall\ p = t - B^T$ to $t - 1$, but Eq. (C.3), which is used to generate the network output is applicable only when p = t – 1.

- Evaluation of activated output of ith node in first hidden layer at time step p $\left(x_i^{1,p}\right)$:

$$y_i^{1,p} = w_{i1}^1\, X^p + \sum_{k=1}^{N^m}\left(\overline{w}_{ik}^1\, x_k^{1,p-1}\right) + b_i^1 \text{ and } x_i^{1,p} = \varphi\left(y_i^{1,p}\right) \quad \text{(C.1)}$$

Here, $y_i^{1,p}$ is the weighted sum of inputs at time step p, $x_i^{1,p-1}$ is the activated output of ith node in first hidden layer at time step p-1 (x is also known as the hidden state), w, $\overline{w}$ and b are feedforward weight, feedback weight and bias, respectively, φ is the activation function and N^m is the number of nodes in first hidden layer.

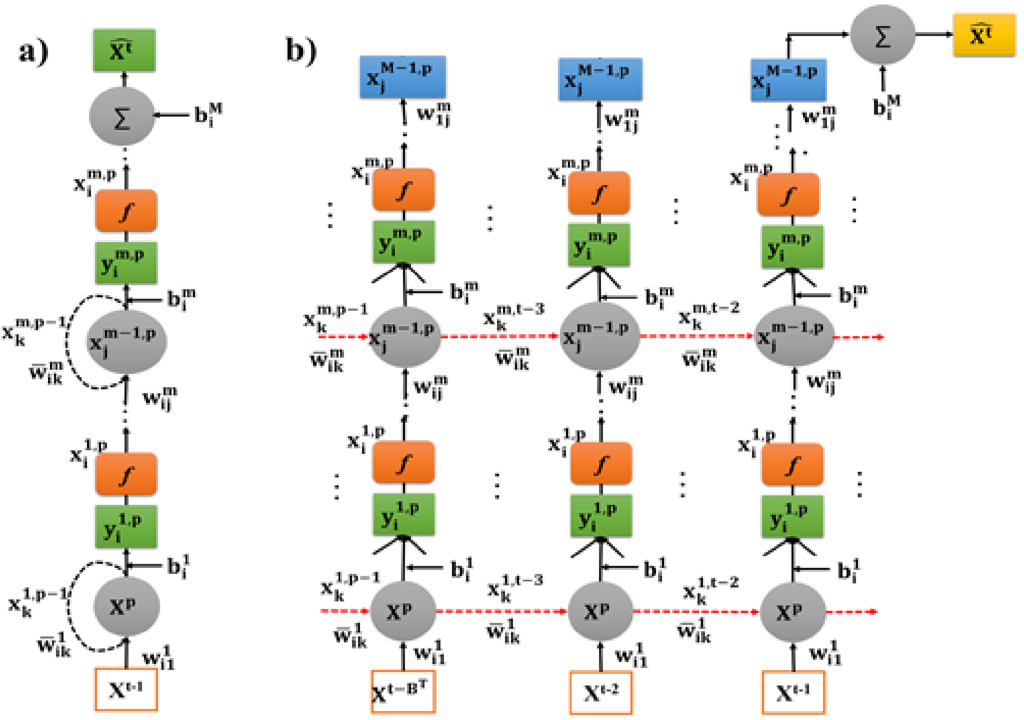


Figure C.1. Pictorial representation of (a) RNN model and (b) unrolled network.

- Evaluation of activated output of i^th node in m^th hidden layer at time step p ($x_i^{m,p}$):

$$y_i^{m,p} = \sum_{j=1}^{N^{m-1}} \left(w_{ij}^m x_j^{m-1,p} \right) + \sum_{k=1}^{N^m} \left(\overline{w}_{ik}^m\, x_k^{m,p-1} \right) + b_i^m \text{ and}$$

$$x_i^{m,p} = \varphi\left(y_i^{m,p}\right) \;\forall\; m = 2 \text{ to } M-1 \qquad \text{(C.2)}$$

- Evaluation of network output ($\hat{X}^t$) only when p = t-1:

$$\hat{X}^t = \sum_{j=1}^{N^{M-1}} \left(w_{1j}^M\, x_j^{M-1,p} \right) + b_1^M \qquad \text{(C.3)}$$

Conventionally, in recurrent networks, output layer does not have a feedback loop. Also, similar to feedforward networks, the activation function in output layer is linear. Thus, the Eq. (C.3) is neither having any feedback connections nor it is having any activation function similar to Eq. (B.5) for WNNs and Eq. (A.3) for NAR models. Further, as opposed to one-to-one style of RNN architecture, where output is evaluated at every time step, the kind of architecture style used in this manuscript (see Figure C.1a) is called many-to-one, where output is evaluated once for every B^T inputs (see Figure C.1b). Since, in this work, the RNNs are compared with WNN and NAR models which consume B^T inputs to generate one output (see Figure A.1 and B.1), many-to-one style of RNN which also requires B^T inputs to generate one output is implemented.

In this manuscript, we use LSTM networks, which belong to the category of deep recurrent neural networks, to model the wind time-series data. This is because, when compared with simple RNNs, the LSTM networks are known to work better when long-term dependencies exist in the dataset. The fact that long-term dependencies are known to be present in a natural time-series data, such as wind speed and direction, is iterated repeatedly in literature and also checked in the current work. For this reason, LSTMs are implemented in the current work instead of simple RNNs.

A node in LSTM network is compared with a node in simple recurrent network in Figure C.2. A primary difference is the additional output from the LSTM node, called the cell state (C), which prevents the problem of vanishing gradients. Based on the context, the LSTM

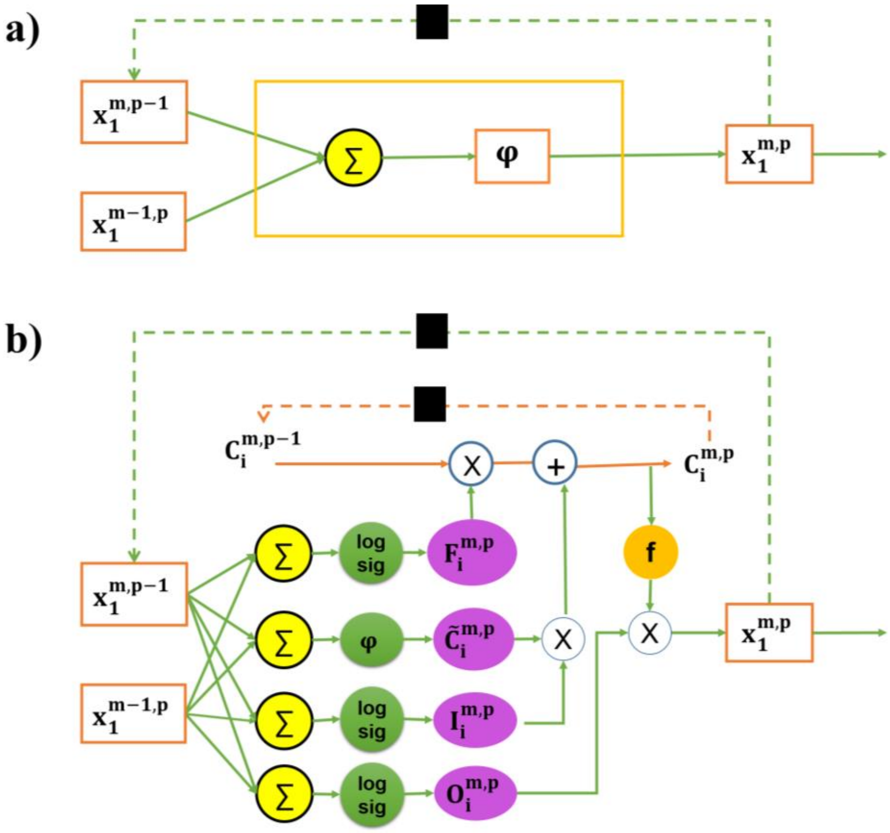


Figure C.2. Comparison between an (a) RNN node and (b) LSTM node.

network is trained to regulate the amount of previous information needed to predict the current output. To facilitate this, LSTM node has four fundamental units described below.

1. Forget gate of ith node in mth hidden layer at time step p $\left(\mathrm{F}_{\mathrm{i}}^{\mathrm{m,p}}\right)$,

$$\mathrm{F}_{\mathrm{i}}^{\mathrm{m,p}} = \mathrm{logsig}\left[\sum_{\mathrm{j=1}}^{\mathrm{N^{m-1}}}\left(\mathrm{wF}_{\mathrm{ij}}^{\mathrm{m}} x_{\mathrm{j}}^{\mathrm{m-1,p}}\right) + \sum_{\mathrm{k=1}}^{\mathrm{N^m}}\left(\overline{\mathrm{wF}}_{\mathrm{ik}}^{\mathrm{m}}\, x_{\mathrm{k}}^{\mathrm{m,p-1}}\right) + \mathrm{bF}_{\mathrm{i}}^{\mathrm{m}}\right] \tag{C.4}$$

2. Input gate of ith node in mth hidden layer at time step p $\left(\mathrm{I}_{\mathrm{i}}^{\mathrm{m,p}}\right)$,

$$\mathrm{I}_{\mathrm{i}}^{\mathrm{m,p}} = \mathrm{logsig}\left[\sum_{\mathrm{j=1}}^{\mathrm{N^{m-1}}}\left(\mathrm{wI}_{\mathrm{ij}}^{\mathrm{m}} x_{\mathrm{j}}^{\mathrm{m-1,p}}\right) + \sum_{\mathrm{k=1}}^{\mathrm{N^m}}\left(\overline{\mathrm{wI}}_{\mathrm{ik}}^{\mathrm{m}}\, x_{\mathrm{k}}^{\mathrm{m,p-1}}\right) + \mathrm{bI}_{\mathrm{i}}^{\mathrm{m}}\right] \tag{C.5}$$

3. Cell of ith node in mth hidden layer at time step p $\left(\tilde{\mathrm{C}}_{\mathrm{i}}^{\mathrm{m,p}}\right)$,

$$\tilde{\mathrm{C}}_{\mathrm{i}}^{\mathrm{m,p}} = \varphi\left[\sum_{\mathrm{j=1}}^{\mathrm{N^{m-1}}}\left(\mathrm{wC}_{\mathrm{ij}}^{\mathrm{m}} x_{\mathrm{j}}^{\mathrm{m-1,p}}\right) + \sum_{\mathrm{k=1}}^{\mathrm{N^m}}\left(\overline{\mathrm{wC}}_{\mathrm{ik}}^{\mathrm{m}}\, x_{\mathrm{k}}^{\mathrm{m,p-1}}\right) + \mathrm{bC}_{\mathrm{i}}^{\mathrm{m}}\right] \tag{C.6}$$

4. Output gate of ith node in mth hidden layer at time step p $\left(\mathrm{O}_{\mathrm{i}}^{\mathrm{m,p}}\right)$,

$$\mathrm{O}_{\mathrm{i}}^{\mathrm{m,p}} = \mathrm{logsig}\left[\sum_{\mathrm{j=1}}^{\mathrm{N^{m-1}}}\left(\mathrm{wO}_{\mathrm{ij}}^{\mathrm{m}} x_{\mathrm{j}}^{\mathrm{m-1,p}}\right) + \sum_{\mathrm{k=1}}^{\mathrm{N^m}}\left(\overline{\mathrm{wO}}_{\mathrm{ik}}^{\mathrm{m}}\, x_{\mathrm{k}}^{\mathrm{m,p-1}}\right) + \mathrm{bO}_{\mathrm{i}}^{\mathrm{m}}\right] \tag{C.7}$$

In these equations, logsig indicates the log-sigmoid function (see Eq. (A.4)), which outputs a real value between 0 and 1; thus the entities in Eq. (C.4), (C.5) and (C.7) are called gates with reference to the logical gates in computer science theory. In contrast with simple RNN node, as mentioned previously, every LSTM node has two outputs – the cell state and the hidden state (which is equivalent to the output of RNN node in Eq. (C.2)). Both these outputs are evaluated using the aforementioned four fundamental units of LSTM node as shown in Eq. (C.8) and (C.9).

- Evaluation of Cell state in ith node in mth hidden layer at time step p $\left(\mathrm{C}_{\mathrm{i}}^{\mathrm{m,p}}\right)$

$$\mathrm{C}_{\mathrm{i}}^{\mathrm{m,p}} = \mathrm{F}_{\mathrm{i}}^{\mathrm{m,p}}\mathrm{C}_{\mathrm{i}}^{\mathrm{m,p-1}} + \mathrm{I}_{\mathrm{i}}^{\mathrm{m,p}}\tilde{\mathrm{C}}_{\mathrm{i}}^{\mathrm{m,p-1}} \tag{C.8}$$

- Evaluation of activated output of ith node in mth hidden layer at time step p ($x_{\mathrm{i}}^{\mathrm{m,p}}$):

$$x_{\mathrm{i}}^{\mathrm{m,p}} = \mathrm{O}_{\mathrm{i}}^{\mathrm{m,p}}\varphi\left(\mathrm{C}_{\mathrm{i}}^{\mathrm{m,p}}\right) \tag{C.9}$$

The auto-regression equation, which has been fundamental to model the time-series is implemented on the cell state in case of LSTMs (see Eq. (C.8)), hidden state in case of RNNs (see Eq. (C.2)), and previous data points in case of WNNs and NAR models (see Eq. (B.3), (B.4) and (A.2)). Following are few important points which are of relevance when LSTMs are considered.

1. As opposed to fixed weights in auto-regression equations in RNNs, WNNS and NAR models, the weights in auto-regression equation in LSTM are the forget and input gates. Since the values of these gates vary with every data point, they regulate the previous and current information needed to evaluate the output of LSTM node at every time point. This allows the LSTM node to have long and short-term memories based on the context in the data.
2. Since the LSTM node provides an auto-regressive output in terms of cell state, which is devoid of any nonlinear activation function (see Eq. (C.8)), it allows the evaluation of gradients (necessary for training) across large length of time-series without vanishing. This allows the LSTM networks to prevent the problem of vanishing gradients.
3. As represented in Figure C.2, except for the difference between the node, rest of the network remains same in case of both simple RNNs and LSTMs. In terms of equations, it means that instead of evaluating the output of node using Eq. (C.2) in RNNs, the same output is evaluated using Eq. (C.4) to Eq. (C.9) in case of LSTMs. Further, if $F = 0$ in Eq. (C.4), $I = 1$ in Eq. (C.5) and $O = 0$ in Eq. (C.7), cell state will be same as the hidden state in simple RNN. Thus, LSTMs enable all the functionalities of simple RNNs. The additional capabilities of LSTM networks, however, come at the cost of additional parameters (about 4 times that of RNN) in them.

Appendix D: List of Pareto solutions obtained for Wind Speed and Wind Direction

Table D.1**:** List of Pareto solutions for wind speed with NAR model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	1	1	0.949	4	-17013
2	1	0	0	1	1	0.949	4	-17013
3	1	0	0	1	1	0.949	4	-17013
4	1	0	0	1	2	0.989	5	-21622
5	1	0	0	1	2	0.989	5	-21622
6	1	0	0	1	2	0.989	5	-21622
7	1	1	0	1	3	0.991	8	-22227
8	1	1	0	2	3	0.991	8	-26476
9	1	1	2	1	3	0.990	13	-22227
10	1	1	4	2	4	0.991	20	-26558
11	1	1	5	1	4	0.991	23	-22305
12	1	2	0	1	5	0.991	13	-22336
13	1	2	1	2	5	0.991	15	-26572
14	1	2	2	1	5	0.991	19	-22326
15	1	2	4	1	6	0.991	28	-22321
16	1	3	0	1	7	0.993	18	-23433
17	1	3	1	2	7	0.994	20	-27669
18	1	3	4	1	8	0.994	36	-23553
19	1	4	0	1	1	0.949	15	-16993
20	1	4	4	2	2	0.989	36	-25813
21	1	4	5	1	2	0.989	42	-21549
22	1	5	0	1	3	0.990	20	-22213
23	2	0	0	1	1	0.949	7	-17008
24	2	3	0	1	7	0.994	29	-23501
25	2	3	1	1	7	0.994	31	-23499
26	**2**	**3**	**4**	**2**	**8**	**0.994**	**48**	**-27979**
27	2	3	7	2	8	0.994	63	-27949
28	3	0	0	1	1	0.949	10	-17002
29	3	0	0	1	1	0.949	10	-17002
30	3	3	2	1	7	0.994	47	-23471
31	3	3	5	2	8	0.994	65	-27952
32	3	3	6	1	8	0.994	70	-23702
33	3	7	1	1	7	0.994	62	-23441
34	3	7	2	1	7	0.994	71	-23429
35	4	0	0	1	1	0.949	13	-16996
36	5	1	0	1	3	0.990	28	-22205
37	5	1	1	1	3	0.990	30	-22203
38	5	1	2	2	3	0.991	33	-26447

Table D.2**:** List of Pareto solutions for wind speed with WNN model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	2	1	0.949	4	-21263
2	1	0	0	2	1	0.949	4	-21263
3	1	0	0	1	2	0.989	6	-25868
4	1	0	0	1	2	0.989	6	-25868
5	1	1	0	1	3	0.990	10	-26466
6	1	2	0	1	5	0.991	17	-26567
7	1	2	1	1	5	0.991	20	-26565
8	1	7	0	1	7	0.993	36	-27419
9	1	7	1	1	7	0.993	44	-27604
10	2	0	0	1	1	0.949	7	-21259
11	2	0	0	1	1	0.949	7	-21259
12	2	0	0	1	1	0.949	7	-21259
13	2	1	4	1	4	0.991	33	-26476
14	2	2	0	1	5	0.991	31	-26555
15	2	3	1	1	7	0.993	48	-27603
16	2	3	2	1	7	0.994	55	-27705
17	**2**	**3**	**4**	**1**	**8**	**0.994**	**73**	**-27765**
18	2	5	4	1	4	0.991	81	-26383
19	3	0	0	1	1	0.949	10	-21253
20	3	0	0	2	1	0.949	10	-21253
21	4	0	0	2	1	0.949	13	-21247
22	4	0	0	2	1	0.949	13	-21247

Table D.3**:** List of Pareto solutions for wind speed with LSTM model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	1	2	0.988	14	-12788
2	1	0	0	2	3	0.989	14	-13537
3	1	0	0	1	10	0.989	14	-13379
4	1	0	0	1	14	0.989	14	-13558
5	1	2	2	1	38	0.993	87	-14112
6	2	0	0	2	3	0.990	35	-13804
7	2	0	0	1	6	0.990	35	-13969
8	2	0	0	1	12	0.991	35	-14142
9	2	0	0	1	16	0.991	35	-14134
10	3	0	0	1	2	0.988	64	-13234
11	3	0	0	2	3	0.990	64	-13658
12	3	0	0	1	4	0.990	64	-13891
13	3	0	0	1	6	0.991	64	-14086
14	3	0	0	1	8	0.992	64	-14594
15	3	0	0	1	12	0.993	64	-15097
16	3	1	0	1	18	0.993	82	-13824
17	3	4	0	2	3	0.990	193	-13368
18	4	0	0	2	3	0.990	101	-13566
19	4	0	0	1	6	0.991	101	-14102
20	4	0	0	2	7	0.992	101	-13382
21	4	0	0	1	8	0.993	101	-14515
22	4	0	0	1	12	0.994	101	-15309
23	4	0	0	1	14	0.995	101	-15557
24	4	1	1	1	20	0.995	134	-14233
25	5	0	0	1	2	0.988	146	-13170
26	5	0	0	2	3	0.990	146	-13575
27	5	0	0	2	7	0.992	146	-14016
28	5	1	2	1	22	0.995	203	-14873
29	**5**	**2**	**0**	**1**	**34**	**0.995**	**207**	**-16103**
30	5	5	0	1	18	0.995	366	-14976
31	6	0	0	2	3	0.990	199	-13325
32	6	0	0	2	7	0.993	199	-14104
33	6	0	0	1	8	0.993	199	-14636
34	7	0	0	1	2	0.988	260	-12941
35	7	0	0	2	7	0.993	260	-14075
36	7	0	0	1	10	0.994	260	-14501
37	7	0	0	1	12	0.994	260	-14792
38	7	0	0	2	13	0.994	260	-13994
39	7	0	0	2	17	0.994	260	-15349
40	7	3	0	1	50	0.995	388	-15703
41	7	4	0	2	3	0.990	449	-12786
42	8	0	0	1	2	0.988	329	-12808
43	8	0	0	2	5	0.991	329	-13461
44	8	0	0	2	7	0.993	329	-14162
45	8	0	0	2	13	0.994	329	-14361

Table D.4**:** List of Pareto solutions for wind Direction with NAR model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	2	1	0.955	4	-20452
2	1	0	0	1	1	0.955	4	-16201
3	1	0	0	1	1	0.955	4	-16201
4	1	0	0	2	1	0.955	4	-20452
5	1	0	0	1	2	0.989	5	-20603
6	1	0	0	1	2	0.989	5	-20603
7	1	0	0	2	2	0.989	5	-24854
8	1	1	0	1	3	0.991	8	-21277
9	1	1	0	2	3	0.991	8	-25526
10	1	1	4	1	4	0.991	20	-21311
11	1	1	4	2	4	0.991	20	-25559
12	1	1	5	1	4	0.991	23	-21305
13	1	2	0	2	5	0.992	13	-25685
14	1	3	0	2	7	0.994	18	-26688
15	1	3	4	2	8	0.995	36	-27062
16	1	4	0	1	1	0.955	15	-16181
17	1	4	2	1	1	0.955	23	-16165
18	1	4	4	1	2	0.989	36	-20546
19	1	4	5	1	2	0.989	42	-20535
20	1	4	6	1	2	0.989	48	-20523
21	1	4	7	1	2	0.989	54	-20511
22	2	1	0	1	3	0.991	13	-21319
23	2	1	0	2	3	0.991	13	-25568
24	2	1	2	2	3	0.991	18	-25560
25	2	1	4	2	4	0.991	26	-25581
26	2	1	5	2	4	0.991	29	-25575
27	2	2	0	1	5	0.992	21	-21443
28	2	2	3	2	5	0.992	31	-25669
29	2	3	0	2	7	0.994	29	-26677
30	2	3	1	2	7	0.994	31	-26684
31	2	3	3	2	7	0.994	41	-26665
32	**2**	**3**	**4**	**2**	**8**	**0.995**	**48**	**-27078**
33	2	4	0	2	1	0.955	21	-20419
34	2	6	0	1	5	0.992	37	-21413
35	3	3	7	2	8	0.995	75	-27037
36	4	3	4	2	8	0.995	72	-27041

Table D.5**:** List of Pareto solutions for wind Direction with WNN model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	1	1	0.955	4	-20452
2	1	0	0	1	1	0.955	4	-20452
3	1	0	0	1	1	0.955	4	-20452
4	1	0	0	1	2	0.989	6	-24831
5	1	0	0	1	2	0.989	6	-24831
6	1	0	0	1	2	0.989	6	-24831
7	1	1	0	1	3	0.991	10	-25523
8	1	1	1	1	3	0.991	12	-25520
9	1	2	0	1	5	0.991	17	-25641
10	1	3	0	1	7	0.994	24	-26641
11	1	3	4	1	8	0.994	51	-26920
12	2	0	0	2	1	0.955	7	-20447
13	2	0	0	2	1	0.955	7	-20447
14	2	0	0	2	2	0.989	11	-24838
15	2	0	0	2	2	0.989	11	-24838
16	2	1	1	1	3	0.991	20	-25511
17	2	1	4	1	4	0.991	33	-25533
18	2	2	0	2	5	0.992	31	-25650
19	2	3	0	1	7	0.994	44	-26656
20	2	3	2	1	7	0.994	55	-26624
21	**2**	**3**	**4**	**1**	**8**	**0.995**	**73**	**-26970**
22	2	3	6	1	8	0.995	87	-26957
23	2	5	0	1	3	0.991	38	-25502
24	3	0	0	2	2	0.989	16	-24846
25	4	0	0	2	2	0.989	21	-24834
26	4	0	0	2	2	0.989	21	-24834
27	4	4	2	1	1	0.955	59	-20344
28	5	0	0	2	1	0.955	16	-20429
29	5	4	0	1	1	0.955	55	-20351
30	7	0	0	1	1	0.955	22	-20417
31	7	0	0	1	1	0.955	22	-20417
32	7	0	0	1	2	0.989	36	-24806
33	7	0	0	1	2	0.989	36	-24806

Table D.6**:** List of Pareto solutions for wind Direction with LSTM model.

S. No.	Number of nodes			Activation Function Choice	Unrolling Length	R^2	Number of parameters	AIC
	Hidden Layer 1	**Hidden Layer 2**	**Hidden Layer 3**					
1	1	0	0	1	2	0.988	14	-13350
2	1	0	0	2	3	0.990	14	-14013
3	1	0	0	1	6	0.990	14	-13954
4	1	0	0	1	12	0.990	14	-13953
5	1	1	0	1	18	0.990	26	-13947
6	**1**	**6**	**0**	**1**	**34**	**0.996**	**211**	**-16606**
7	2	0	0	1	2	0.988	35	-13532
8	2	0	0	2	3	0.991	35	-14156
9	2	0	0	2	5	0.991	35	-14215
10	2	0	0	2	9	0.992	35	-14502
11	3	0	0	1	2	0.989	64	-13575
12	3	0	0	1	4	0.991	64	-14328
13	3	0	0	2	7	0.992	64	-14248
14	3	0	0	2	9	0.993	64	-14859
15	3	0	0	1	12	0.993	64	-14943
16	3	0	0	1	14	0.994	64	-14910
17	3	2	0	1	34	0.995	111	-14897
18	3	6	0	1	34	0.996	307	-16524
19	4	0	0	2	3	0.991	101	-14066
20	4	0	0	1	4	0.991	101	-14289
21	4	0	0	2	5	0.991	101	-14294
22	4	0	0	2	7	0.993	101	-14662
23	4	0	0	1	8	0.994	101	-15429
24	4	0	0	1	10	0.994	101	-15334
25	4	0	0	1	12	0.995	101	-15888
26	4	0	0	1	14	0.995	101	-16035
27	4	0	0	1	16	0.995	101	-16251
28	5	0	0	1	2	0.989	146	-13300
29	5	0	0	1	6	0.991	146	-14230
30	5	0	0	2	7	0.993	146	-14138
31	5	0	0	1	8	0.994	146	-15434
32	5	0	0	2	15	0.995	146	-15028
33	5	1	0	1	18	0.995	170	-15887
34	5	2	0	1	34	0.995	207	-16365
35	6	0	0	2	5	0.991	199	-14108
36	6	0	0	1	10	0.994	199	-15442
37	6	0	0	2	15	0.995	199	-15524
38	6	1	0	1	18	0.995	226	-16081
39	7	0	0	2	3	0.991	260	-13829
40	7	0	0	1	4	0.991	260	-14015
41	7	0	0	2	7	0.993	260	-14360
42	7	0	0	1	10	0.994	260	-15338
43	7	0	0	1	14	0.995	260	-15885
44	7	1	1	1	20	0.995	302	-15428
45	8	0	0	2	5	0.991	329	-13821
46	8	0	0	2	7	0.993	329	-14426
47	8	0	0	2	9	0.994	329	-14782
48	8	1	0	1	18	0.995	362	-15779

Highlights:

- Time series Analysis and Decomposition of wind time series data.
- Usage of real wind characteristics data to validate the proposed algorithm.
- Significance of long-term variability for accurate forecasting of Wind is studied.
- Evolutionary Neural Architecture Search inspired by Green Deep Learning.
- Comprehensive Comparison of optimal LSTMs, optimal NAR and optimal WNNs.

ORCID Information:

Kishalay Mitra

Indian Institute of Technology Hyderabad, Kandi, Telangana, India

kishalay@che.iith.ac.in

ORCID ID: 0000-0001-5660-6878

Credit Author Statement

Pujari, K.N.: Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing - original draft.

Miriyala, S.S.: Conceptualization; Methodology; Supervision; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft, review & editing.

Mittal, P.: Conceptualization; Formal analysis; Supervision; Software; Visualization.

Mitra, K.: Conceptualization; Formal analysis; Funding acquisition; Methodology; Project administration; Resources; Software; Supervision; Visualization; Writing - review & editing.

Declaration of Interest Statement

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: