J.Natn.Sci.Foundation Sri Lanka 2022 50 (2): 503 - 517
DOI: http://dx.doi.org/10.4038/jnsfsr.v50i2.10591

RESEARCH ARTICLE

Wind energy prediction

Development of wind energy prediction models using statistical, machine learning and hybrid techniques: a case study

P Ekanayake[1], O Panahatipola[1] and J Jayasinghe[2*]
[1] *Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka.*
[2] *Department of Electronics, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka.*

Submitted: 21 June 2021; Revised: 18 January 2022; Accepted: 25 March 2022

Abstract: This paper presents the development of statistical, machine learning, and hybrid models to predict the wind energy generation of a major wind farm in Sri Lanka named Nala Danavi. Regression based statistical techniques, namely, Multiple Linear Regression and Power Regression were applied to the input variables of wind speed and ambient temperature producing expressions for wind energy in terms of those weather indices. Similarly, the machine learning techniques of Support Vector Regression and Artificial Neural Network were applied in developing another set of wind energy prediction models. Unlike in the methods mentioned above, the Support Vector Machine was applied only to the past energy data. In contrast, a Vector Error Correction model was developed by using both wind energy and weather data. Further, time series modelling of wind energy data was performed to develop a Seasonal Autoregressive Integrated Moving Average model. The research was further extended to develop two hybrid prediction models by combining Support Vector Machine and Artificial Neural Network each with Seasonal Autoregressive Integrated Moving Average. The performance of all the models was assessed and compared in terms of the Coefficient of Determination, Root Mean Square Error, and the Mean Absolute Error. As per the results, all the models were highly accurate while the Support Vector Regression produced the most precise prediction model. The models developed in this research can be used to predict the wind energy generated at Nala Danavi wind farm using either the previous energy data or the projected weather data of the region.

Keywords: Correlation, neural networks, regression analysis, time series analysis, wind energy.

INTRODUCTION

Among all basic needs of human life today, the availability of an uninterrupted power supply has become indispensable due to the overwhelming dependence of every economic, educational, and social activity of people on electricity. The demand for electricity shall continue to increase in proportion to the growth of world population and other factors like urbanization and industrialization. With strong emphasis on ensuring a decarbonized environment for the future generation, the global energy bodies like the International Energy Agency (IEA) and the Global Wind Energy Council (GWEC) are insisting on the adoption of environmentally friendly renewable energy sources, and discouraging the use of power sources like thermal and fossil fuel that cause air pollution and ozone depletion leading to global warming. According to the World Energy Outlook (2018) by the IEA, the worldwide electricity consumption would increase by more than 25% within the next two decades. In this context, the environmental security will depend on the extent or the contribution made by clean energy sources like solar and wind power. Of these, the Global Wind Report (2019) by the GWEC reports that global wind power installations are gathering momentum with 60.4 GW of new input in 2019 alone due to its cost effectiveness compared to other renewable sources like bioenergy, tidal energy, hydro energy and

[*] Corresponding author (Jeevani@wyb.ac.lk; https://orcid.org/0000-0002-9266-8643)

This article is published under the Creative Commons CC-BY-ND License (http://creativecommons.org/licenses/by-nd/4.0/). This license permits use, distribution and reproduction, commercial and non-commercial, provided that the original work is properly cited and is not changed in anyway.

solar photovoltaic energy. The annual increment of wind power capacity is reported to be around 30%, which has dominated the growth of other renewable energy sources during the past decade (Wang *et al.*, 2011; Zhao *et al.*, 2016). However, there is no shortage of challenges in extracting wind power to the main power grid from both onshore and offshore wind installations, due to the natural instability of weather around them.

The integration of wind power into a main grid also requires extra planning and management for balancing the power systems in the presence of the intermittence and volatility of wind power over various time scales (Foley *et al.*, 2010). The greatest impediment in the process of wind power assimilation to a main power grid is caused by fluctuations and uncertainties in wind speeds (Hanifi *et al.*, 2020), which set the stage for extensive research on discovering accurate wind power forecasting methods based on physical, statistical, machine learning and more recently hybrid techniques (Liu *et al.*, 2010; Shi *et al.*, 2012; Hanifi *et al.*, 2020). It has been pointed out that the risk of unpredictability in wind power penetration can be minimized with precise forecasting of wind speed and power generation (Chang, 2014). Wind speed is known to be affected by temperature, direction of wind, humidity, etc., as well as topographical and temporal features like terrain and particular time of the day/season. The demand for wind power forecasting over broad time scales is greater due to its advantages in planning unit commitment, dispatching, maintenance scheduling, and maximizing profits in power trade (Santhosh *et al.*, 2020).

The statistical models used in wind power forecasting have both advantages and disadvantages. Economical usage with convenience, accuracy in predictions over short-time scales (up to one day ahead), adaptation to the prevailing situation using auto-recursive mathematical algorithms, and the ability to provide average or peak wind speeds over such time scales are reported to be the advantages, while inaccurate predictions over long time periods (month or more ahead), need of a testing and tuning process to generate accurate predictions, and the assistance of site experts are reported to be the disadvantages (Lehr & Keeley, 2016). The Auto Regressive Moving Average (ARMA) is a widely used statistical model based on time series analysis (Chandra *et al.*, 2013). ARMA models, after being applied to wind speed and wind power output, have also shown that statistical modelling greatly improves wind forecasts compared to persistent forecasts (Milligan *et al.*, 2004). Taking the non-stationary wind speed into account and using a Kalman Filter to estimate the time-varying model parameters, a linear autoregressive model was developed (Huang & Chalabi, 1995). Defining a particular state with the variables of wind speed, direction, and wind power, and determining the transition matrix using a maximum likelihood estimator based on multi-step transition data collected from a wind turbine in Portugal, a Markov Chain model also explained the wind speed and power probability density (Lopes *et al.*, 2012).

With the developments in the field of data science, researchers started to explore the potential of developing models based on machine learning for the prediction of wind energy generation and the performance of those models were compared with traditional statistical models. Having applied time series analysis on mean hourly wind speed data, a comparative study of several forecasting methods was conducted with ARMA, Feed Forward and Recurrent Neutral Networks, Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Neural Logic Networks (NLN) (Sfetsos, 2000). As per the research findings of that study, neural network-based models outperformed ARMA. Yona *et al.* (2013) used historical wind speed data in a time series analysis using Neural Network, Auto Regression, and Kalman Filter to test wind power generation at a target time and their models predicted the power output based on wind forecasts. Using past power data and forecasts of wind speed and direction at a wind farm, a combination of Neural Networks and fuzzy logic techniques have also been applied to predict the wind power output (Sideratos & Hatziargyriou, 2007). In a review on different methods of wind power prediction, Agarwal *et al.* (2018) categorized statistical approaches into methods based on Neutral Networks and models depending on time series. Though not generalized, they further pointed out that the performance of Neural Networks is usually higher than that of the time series models for most time scales. In another study, an ARMA statistical model and an Artificial Neural Network (ANN) model based on the Feed Forward Neutral Network were used to forecast the wind power generation in Suthari, India while measuring the model accuracy in terms of the Mean Absolute Percentage Error (MAPE) and Mean Square Error (MSE) (Bhatt & Gandhi, 2019). Having used the normalization technique for better training performance and using the correlation coefficient to select inputs to the model, the ANN results were reported to be remarkably better than those produced by the statistical model. As per the literature, there is a high probability for techniques based on machine learning to outperform statistical approaches in the prediction of wind energy generation.

Apart from using statistical or machine learning techniques in isolation, several recent studies in the literature have highlighted the plausible dexterity of hybrid models for the prediction of wind power over conventional techniques (Shi *et al*., 2012). A comprehensive study was carried out by Shi *et al.* (2012) to predict the hourly wind speed and power generation for five forecasting horizons using the hybrid methodology. In their work, the linear component of time series data was modelled by the ARIMA method and the nonlinear component by the ANN and Support Vector Machine (SVM) methods, to make up the two hybrid models: ARIMA-ANN and ARIMA-SVM. Having compared the results generated by the three single and two hybrid models, it was shown that the overall performance difference across all forecasting horizons remained under 5.5%. Due to this marginality of results, they concluded that hybrid models do not always yield the best predictions. Perusing the literature, it could be found that hybrid models were scarcely used for the wind energy prediction. In the meantime, hybrid modelling has been used in other fields too. Tseng *et al.* (2002) reported the development of a hybrid prediction model that comprised of Seasonal ARIMA (SARIMA) and the Back Propagation Neural Network models. Having combined the statistical ARIMA model with two Neural Networks, it was shown in a separate study that the hybrid model produced better results over each model applied separately (Hansen & Nelson, 2003). Voyant *et al.* (2012) used a pre-input layer selection method and combined an optimized multilayer perception with an ARMA model. They proved the reduction in the prediction error of the ANN-ARMA hybrid method and pointed out that their conclusions are consistent with renewable energy deployment too. In this context, this research focuses on the development of conventional statistical and machine learning models as well as hybrid models for the wind energy prediction.

Though the generation of wind power is increasing progressively in Sri Lanka, research on the operation of wind farms is in its infancy. Being a major generator of wind energy in Sri Lanka, the Nala Danavi wind farm should possess regular and accurate estimates of power generation capacity in order to manage its wind power contribution to the national grid. However, such estimates are complicated by volatile weather conditions in the region caused by changing wind speeds and other weather indices, which lead to inaccurate predictions. In this study, a comprehensive modelling package is developed to provide precise predictions of wind energy based on either past wind energy and weather data or projected weather data for the future, so that wind power operators can make use of several options. As a previous study has not been undertaken for Nala Danavi with multiple modelling techniques, forecasting of wind energy generation at Nala Danavi wind farm in Sri Lanka is explored in this study comparatively using four statistical models: Multiple Linear Regression (MLR); Power Regression (PR); SARIMA; and Vector Error Correction (VEC), three machine learning models: Support Vector Regression (SVR); SVM; and ANN, and two hybrid models: SARIMA-SVR and SARIMA-ANN. Exploration of hybrid models shall be beneficial in determining whether they are in fact superior for predictions over the use of other techniques for coastal wind installations in Sri Lanka. The models to be identified as appropriate tools for Nala Danavi may be extended to other coastal wind farms in the island that exhibit similar weather patterns over longer periods of time.

MATERIALS AND METHODS

Study area and climatic data

Study area

The Nala Danavi wind farm is located in the Kalpitiya Peninsula in the Puttalam district (08°05′23″N 79°42′33″E) on the western coast of Sri Lanka (Figure 1a). The facility comprises of six turbines (model: Gamesa G58) with the rated power of 850 KW in each. The power curve of the wind turbines is shown in Figure 1b. The cut-in and cut-out wind speeds of the wind turbine are 3.0 m/s and 20.0 m/s respectively. The variation of monthly wind energy output during the period 2015–2020 is shown in Figure 2. The average monthly wind energy generation of the farm is 1112.9 MWh with minimum and maximum monthly energy generation of 125.2 MWh and 3037.8 MWh respectively. Further, an annual pattern of energy generation could be observed with a peak in the middle of each year (Figure 3a).

Wind energy generation zoomed into the daily values is illustrated in Figure 3b. The daily generation in most of the days from May to September varies between 50–100 MWh while it is lower than 40 MWh in other months. On average, about 77% of the annual energy generation is achieved during the period of the above five months as per the past data.

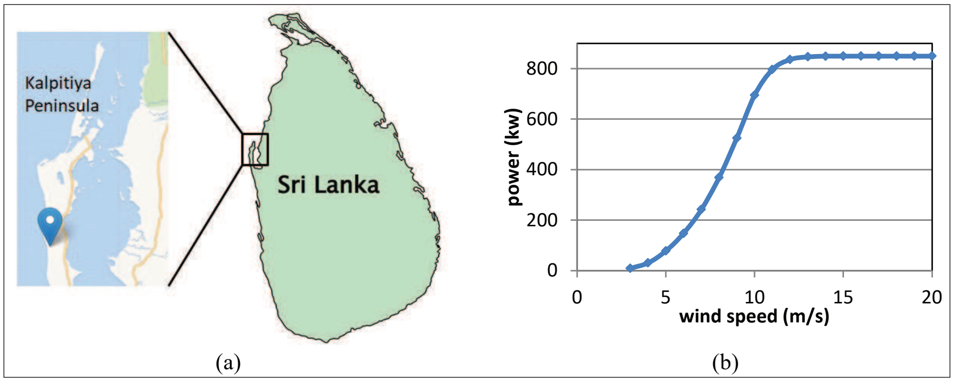

Figure 1: Wind turbines (a) location; (b) power curve

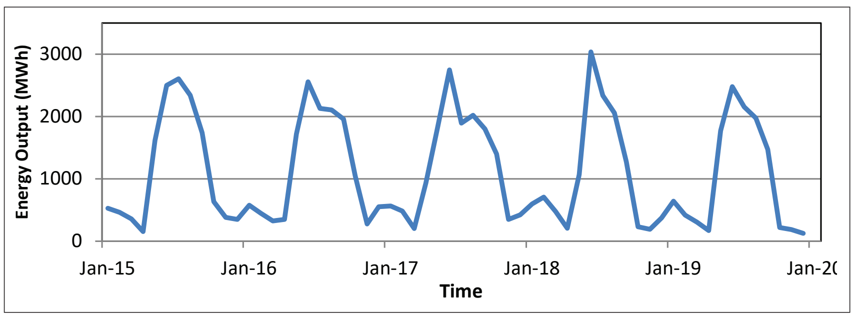

Figure 2: Variation of monthly wind energy output in the period: 2015–2020

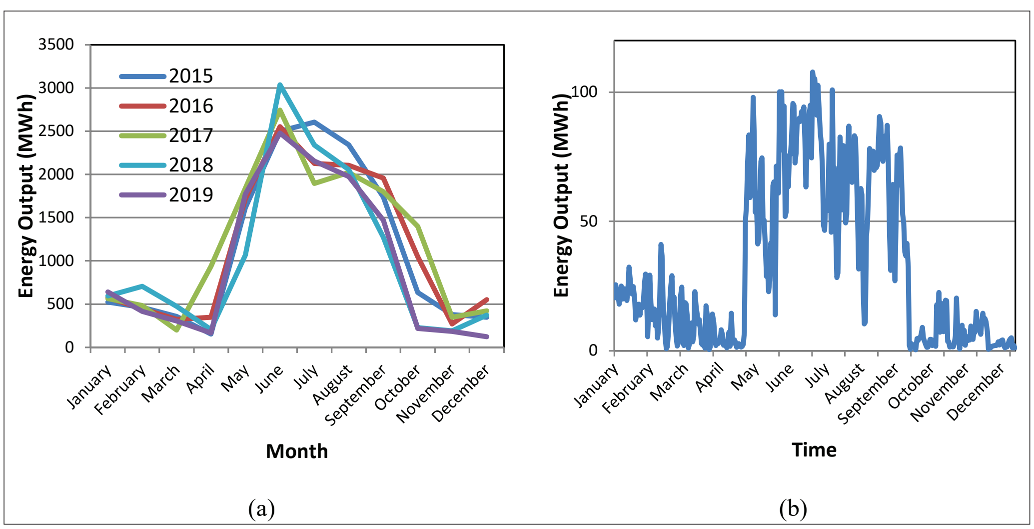

Figure 3: Variation of wind energy output (a) monthly data; (b) daily data

Climatic data

The climatic data of wind speed, wind direction, and the ambient temperature measured at the meteorological station located at the wind farm were obtained. The variation of monthly wind speed (Figure 4a) is similar to the variation of monthly energy generation, having speedy winds during the period from May to September annually. The average monthly wind speed varies between 10.7 m/s and 2.5 m/s, while the instantaneous wind speed varies between 0–18.7 m/s.

As per the plot of wind energy output with wind speed (Figure 5a), they are linearly correlated following the power curve of the wind turbines when the wind speed lies between 3–12 m/s. The monthly ambient temperature varies between 33.7 ºC and 42.4 ºC (Figure 4b). However, as per the data collected throughout the day, the ambient temperature varies between 25ºC and 53 ºC. As per the plot of wind energy output with ambient temperature (Figure 5b), the relationship is not very linear.

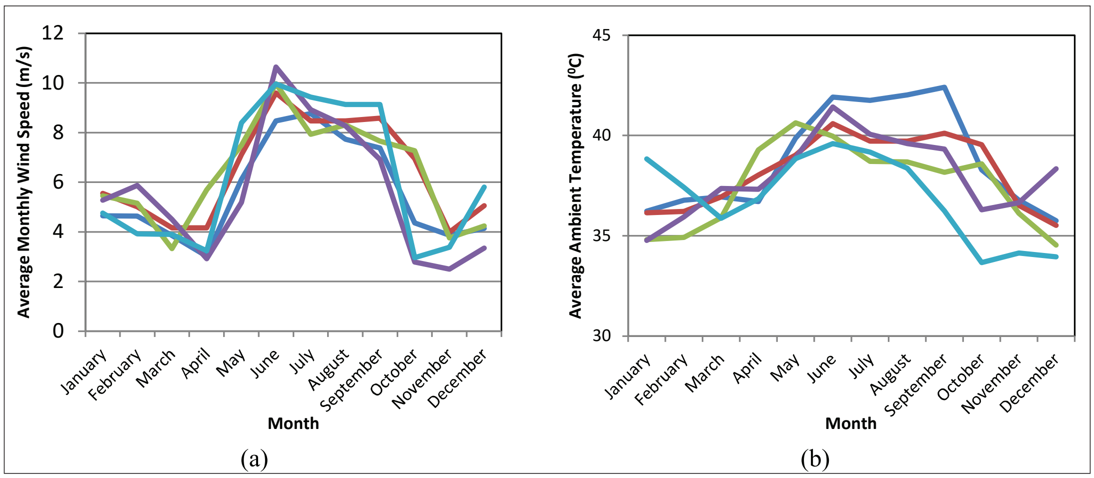

Figure 4: Variation of monthly climatic data (a) wind speed; (b) ambient temperature

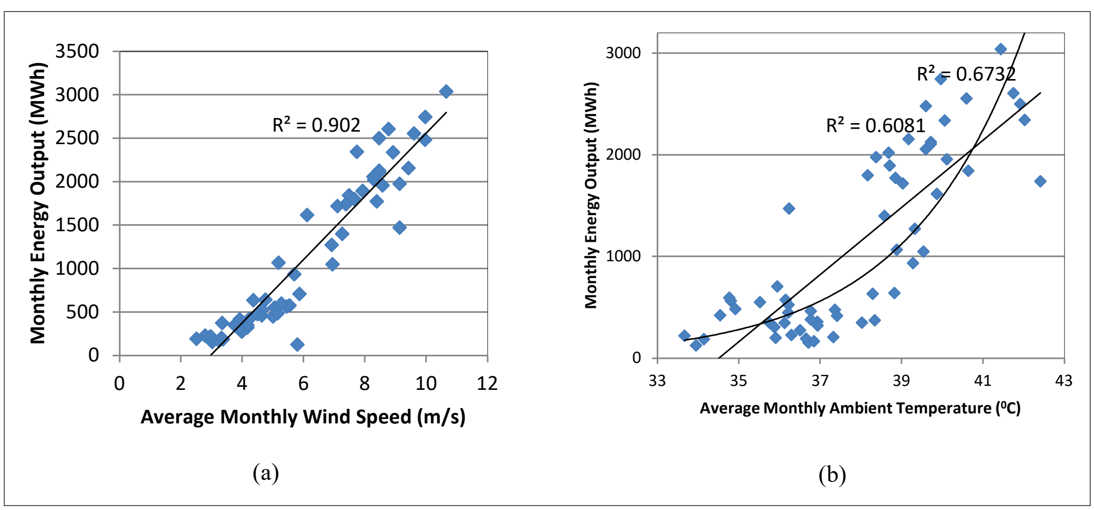

Figure 5: Variation of wind energy output with climatic factors (a) monthly wind speed; (b) monthly ambient temperature

Collinearity

The correlations between the variables were calculated in terms of the Pearson's correlation (R) and the Spearman's correlation (ρ), summarized in Figure 6.

It can be understood from the correlation values that a strong positive correlation exists between wind energy and wind speed. Further, the mediocre correlation values between wind energy and ambient temperature are in agreement with the coherence between the patterns of monthly variation of the same weather indices seen in Figure 4. The wind energy generation depends on the wind direction as well. However, the rotor of the wind turbine rotates continuously based on the wind direction to harvest the maximum wind. Therefore, the wind direction was not considered as an input variable in this study.

	Wind Energy	Ambient Temperature	Wind Speed
Wind energy	1		
Ambient temperature	R= 0.82 ρ = 0.79	1	
Wind speed	R= 0.95 ρ = 0.94	R= 0.69 ρ = 0.67	1

Figure 6: Correlation between the wind energy and weather parameters

Prediction models

Multiple linear regression (MLR) model

The wind energy in MLR is expressed in terms of the wind speed (WS) and temperature (T) as follows:

$$\text{Energy} = \beta_0 + \beta_1 WS + \beta_2 T + \varepsilon \qquad ...(1)$$

where, β_0 is the intercept (a constant), β_1 and β_2 are the regression coefficients of the input variables, and ε is the random error under the assumption that it is normally distributed with mean zero and constant variance.

Power regression (PR) model

Unlike MLR, PR forms a nonlinear relationship in which the output is modelled in proportion to a power of the explanatory variables. Its expression is a power (polynomial) equation of the form of $y = ax^b$, where x has to be non-zero as follows:

$$\text{Energy} = a\, WS^b\ T^c \qquad ...(2)$$

where, *a*, *b*, and *c* are constants.

Seasonal autoregressive integrated moving average (SARIMA) model

ARIMA is one of the most widely used and recognized statistical forecasting time series models. As the wind energy data demonstrate a seasonal pattern with a frequency of one year, a modified version of the ARIMA model called the SARIMA model was used in this research. The structure of the SARIMA model is presented in the form of SARIMA (p, d, q) $(P, D, Q)_s$, where, p, d, and q are non-seasonal terms and P, D, and Q are seasonal terms, which represent the order of the autoregressive process (P and p), the number of differencing (D and d), and the order of the moving average process (Q and q) respectively. The span of seasonality is indicated by s, which is 12 in this research study due to the use of monthly energy data with an annual pattern. The SARIMA model is expressed as (Haddad *et al.*, 2019),

$$\phi_p(B)\Phi_P(B^s)W_t = \theta_q(B)\Theta_Q(B^s)Z_t \qquad ...(3)$$

where, Z_t is a purely random process and W_t is the transformation of the input variable (wind energy) X_t according to

$$W_t = \nabla_s^D \nabla^d X_t. \qquad ...(4)$$

Further, the model terms are expanded as follows.

$$\begin{aligned} \phi_p(B) &= 1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p \\ \Phi_P(B) &= 1 - \alpha_1 B^s - \alpha_2 B^{2s} - \cdots - \alpha_P B^{Ps} \\ \theta_q(B) &= 1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q \\ \Theta_Q(B) &= 1 + \beta_1 B^s + \beta_2 B^{2s} + \cdots + \beta_Q B^{Qs} \end{aligned} \qquad ...(5)$$

In the development of the SARIMA model, the Box-Jenkins Method (Box & Jenkins, 1976) was used for identifying, estimating, and checking models. In the model identification process, several models were found suitable for time series modelling. The auto correlation function (ACF) and partial autocorrelation function (PACF) were applied to make the first guess on the orders (p, q, P, Q) of the ARIMA model, which is the initial tentative model. Using possible combinations of the orders, a set of suitable models was developed. Out of these models (combination of orders for the autoregressive and moving average), the model that generated the minimum Bayesian Information Criterion (BIC) value was selected for further analysis. The BIC is given by,

$$\mathrm{BIC} = \ln(\widehat{\sigma}_e^2) + \frac{2k\ln(n)}{n} \quad ...(6)$$

where, n is the sample size, k is the number of estimated parameters, and $\widehat{\sigma}_e^2$ is the error variance. A major advantage of this criterion is that it can effectively address the problem of overfitting.

Once the model is identified, the parameters need to be estimated and in principle the selected parameters should generate the lowest residual. This can be accomplished by using the Yule–Walker Estimation or the Maximum Likelihood Estimation. After estimating the parameters, the residuals should be checked to assess whether they are random and uncorrelated.

In order to test the selected model, the correlogram of the residuals was analysed and a global contrast was applied using the Ljung–Box statistic mentioned below.

$$Q^* = n\sum_{k=1}^{L} r_k^2(a) \quad ...(7)$$

where, $r_k^2(a)$ is the autocorrelation coefficient of the residuals, *n* is the number of data points and *L* is the maximum delay considered (Shi *et al.*, 2012). In this study, all the data points were used to develop the model and 20% of data points were randomly selected for validation.

Based on the above technique, the data were tested using the Augmented Dickey Fuller Test and it was found that the data are stationary at the 5% significance level. Figure 7 shows the autocorrelation and partial autocorrelation values of the variable, wind energy, by k time periods apart called the lag.

As per the ACF and PACF plots (Figure 7), a model with the structure of SARIMA $(2, 0, 1)(0, 0, 1)_{12}$ was identified tentatively. The set of possible candidate models identified by modifying the tentative model is given in Table 1. The smallest BIC value could be obtained when the structure of the model is in the form of SARIMA $(0, 0, 0)(0, 1, 1)_{12}$.

The corresponding model can be given in the form of

$$\hat{X}_t = a + X_{t-12} + bZ_{t-12} \quad ...(8)$$

where, *a* is a constant and *b* is the seasonal moving average parameter. In order to test the goodness of fit of the proposed SARIMA model, the correlogram of the residuals and Ljung–Box statistics were analyzed. According to the results, the residuals of the proposed model are not correlated.

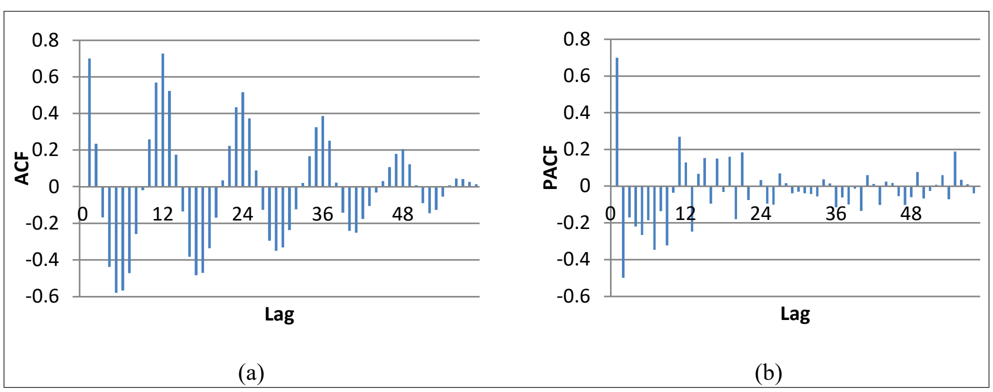

Figure 7: Correlation structure of wind energy (a) auto correlation function; (ACF) (b) partial autocorrelation function (PACF)

Table 1: BIC values of possible candidate models

Model structure	BIC
SARIMA(2,0,1)(0,0,1)$_{12}$	12.4123
SARIMA(1,0,1)(0,0,1)$_{12}$	12.5122
SARIMA(2,0,0)(0,0,1)$_{12}$	12.4492
SARIMA(0,0,0)(0,1,1)$_{12}$	11.4758
SARIMA(0,0,0)(1,1,0)$_{12}$	11.8334

Vector error correction (VEC) model

A VEC Model, which is a special case of vector autoregression, was also used to predict the wind energy. It takes the form of

$$\Delta y_t = \alpha + \sum_{i=1}^{p} \beta_i \, \Delta y_{t-i} + \sum_{i=1}^{p} \gamma_i \, \Delta x_{t-i} + \sum_{i=1}^{p} \delta_i \, \Delta z_{t-i} + \sum_{i=1}^{r} \theta_i \, ECT_i + \varepsilon_t \qquad ...(9)$$

where, Δy_t is the first difference of the variable y_t (wind energy), α is a constant, β is the coefficient of the lags of differenced variable y_t, γ is the coefficient of the lags of differenced variable x (wind Speed), δ is the coefficient of the lags of differenced variable z (temperature), θ is the vector of cointegration relationships, ECT refers to the error correction terms derived from long run cointegration relationships, and ε_t are the serially uncorrelated random error terms with zero mean. Further, the summations are taken over the number of lags p and the rank of the cointegration (r).

80% of data were used to build the model and the remainder was used to validate the model. The Dickey-Fuller unit root test was used to check the stationarity of each variable. Lag order was determined by using the Akaike Information Criterion (AIC), Hannan–Quinn information Criterion (HQC), Schwarz Criterion (SC), and the Final Prediction Error (FPE) information criterion. A cointegration test was used to check whether there is a correlation between several time series in the long term. Further, the Johansen test was used to identify the rank of the cointegration. Seasonal variation was controlled by specifying the season as 12 since data demonstrated a seasonal variation.

After fitting the model, its adequacy was checked in terms of the serial correlation of residuals, heteroscedasticity in the residuals, and the normality of residuals by performing Portmanteau Test, Lagrange-Multiplier test, and Jarque–Bera test respectively. Data were examined for stationarity using the Augmented Dickey Fuller test, which suggested that all the variables are stationary at 5% level of significance. The order of the model was chosen according to the information criteria and three out of four criteria favoured choosing the lag order 2, with AIC(n)=2, HQ(n) = 2, SC(n) = 1, and FPE(n) = 2. The trace of the Johansen test identified the rank of the cointegration as 1.

The model was built choosing lag p = 2 and the rank of cointegration r = 1 by specifying 12 seasonal dummies. After fitting the model, its adequacy was checked, and the results of the Portmanteau Test suggested that the error term is not serially correlated. Further, there was no heteroscedasticity as per the Auto Regressive Conditional Heteroscedasticity (ARCH) Lagrange-Multiplier test. The Jarque–Bera test suggested the normality of the residuals.

Support vector machine (SVM) and support vector regression (SVR) models

SVM is a versatile Supervised Learning Algorithm, which is used for classification as well as regression problems. SVR finds a linear hyper plane, which fits the multidimensional input vectors to output values. In this research, a single SVM model was used to classify monthly wind energy data, and the SVR predicts the monthly wind energy by regressing it on wind speed and temperature, using the support vector algorithms. The standard SVM to solve approximation problems is as follows,

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) k(x_i, x) + b \qquad ...(10)$$

where, α_i^* and α_i are Lagrange multipliers and b is the displacement of bias. The kernel function $k(x_i, x)$ is the inner product of two elements of Hilbert space such that,

$$\mathrm{k}(\mathrm{x_i}, \mathrm{x}) = \varphi(\mathrm{x_i})\varphi(\mathrm{x}). \qquad ...(11)$$

where φ is the nonlinear mapping.
The coefficients α_i^* and α_i have been obtained by minimizing the following regularized risk function:

$$R_{reg}[f] = \frac{1}{2}||\omega||^2 + C \sum_{i=1}^{l} L_\varepsilon(y) \qquad ...(12)$$

The term $||\omega||^2$ has been characterized as model complexity, C is a constant determining the trade-off, ε is the intensive loss function and $L_\varepsilon(y)$ is given by (Li *et al.*, 2020)

$$L_\varepsilon(y) = \begin{cases} 0, & for\ |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & otherwise. \end{cases} \quad ...(13)$$

The kernel function used in this research is the Radial Basis Function (RBF),

$$k(x_i, x) = \exp\left(-\gamma||x_i - x||^2\right)\ ; \qquad \gamma = {}^{1}/_{2\sigma^2}\ . \quad ...(14)$$

The RBF kernel function was selected since the model trained by the RBF has a better overall performance than other kernel functions when there is no prior knowledge about the problems. In the SVM-based prediction models, the constant determining the trade-off (C) and kernel function parameter γ have a great effect on prediction results. The kernel function parameter γ is used to control the model complexity and the approximation error (Li *et al.*, 2020). A grid-search was done using 10-fold cross validation and the optimum values of C, γ and ε were selected based on the MSE. The grid selections of γ, C and ε were (1, 2, 3, 4, 5, ……, 100), (2^{-3}, 2^{-2}, 2^{-1}) and (10^{-4}, 10^{-3}, 10^{-2}) respectively. In this case too, 80% of data from the whole data set was extracted randomly to train the model and the remaining 20% was used to validate the model. Using the grid search, the final optimal parameters which minimized the MSE were found as C = 95, γ = 0.125 and ε = 10^{-4} for SVM and C = 7, γ = 0.125 and ε = 10^{-4} for SVR.

Artificial neural network (ANN)

The Feed-forward Neural Network architecture was trained by using the "neuralnet" function of the R package (Version 4.0.5), which is a network training function that updates weights and bias values during training. The input vector is presented by $Y_j = \{y_1, y_2, y_3, \ldots, . y_n\}$; $W_{jk}(j = 1,2,3, \ldots, n; k = 1,2,3, \ldots, m)$ is the connection weight vector of the *j* nodes of the input layer to the *k* nodes of the hidden layer; $X_k(k = 1,2,3, \ldots, m)$ is the vector of *k* neurons in the hidden layer; $W_k(k = 1,2,3, \ldots, m)$ is the connection weights of the *k* nodes of the hidden layer to the output layer; and *Y* is the unit output vector for the neural network with one output neuron. $\Theta_k(k = 1,2,3,,,,, m)$ is the bias value of the hidden layer nodes and Θ is the bias value of the output layer.

The output of the hidden layer is determined by the formula

$$X_k = f(\textstyle\sum_{j=1}^{n} W_{jk} Y_j + \Theta_k) \quad ...(15)$$

The output of the output layer is determined by the formula (Mapuwei *et al.*, 2020)

$$Y = f(\textstyle\sum_{k=1}^{m} W_k X_k + \Theta), \quad ...(16)$$

where, *f* is the activation function for the hidden and output layers. Here, Logistic (Sigmoid) function was used as the activation function. When designing ANN, data were preprocessed before modelling. The data were normalized using the min-max normalization method in order to prevent saturation of hidden nodes before feeding into the neural network. Data were separated into two sets having 80% of data as the training set and the rest for testing. Supervised training with resilient backpropagation was used as the training algorithm with the threshold value of 0.01 for the training dataset.

Hybrid models

The performances of some forecasting methods such as ARIMA depend upon the linearity of data. Hybrid models, which have the capability to capture both linear and nonlinear characteristics, are thus believed to be a good strategy for time series forecasting. Hybrid models typically employ an ARIMA to model the linearity of data, while using an ANN or SVM for modelling the nonlinearity. In general, the concept of a hybrid model can be represented as follows (Shi *et al.*, 2012):

$$y_t = L_t + N_t, \quad ...(17)$$

where, L_t denotes the linear component and N_t, represents the non-linear component. First an ARIMA model was used to capture the linearity of the data and the non-linear techniques of SVM and ANN were used to capture the nonlinearity of the data.

In the SARIMA-SVM model, the SVM model was constructed to forecast the residuals of the established SARIMA models, and the parameter selection was done as per the same procedure in the single SVM modelling approach. Similarly, in the SARIMA-ANN model, ANN model dealt with the residuals and the FFNN model was developed using the residuals from the trained set, which is 80% of the residuals, and compared against observed data for the remaining test set. In the grid search of the SVM model for the errors established from SARIMA, the final optimal parameters which minimized the MSE were found as C = 2, γ = 0.125 and ε = 10^{-4}. The SARIMA-ANN hybrid model was developed using 5 hidden nodes.

Performance evaluation criteria

The coefficient of determination (R^2) describes the degree of collinearity between observed and predicted data. It indicates the proportion of the variance in observed data explained by the model. In the range from 0 to 1, higher R^2 values indicate less error variance. Typically, a prediction model with $R^2 \geq 0.5$ is considered acceptable (Mapuwei *et al.*, 2020). Although widely used for the evaluation of prediction models, it is oversensitive to outliers. On the other hand, it is insensitive to additive and proportional differences between the observed and predicted data.

$$R^2 = \frac{(\sum_{i=1}^{N}(O_i-\bar{O})(P_i-P_{mean}))^2}{\sum_{i=1}^{N}(O_i-\bar{P})^2 \sum_{i=1}^{N}(P_i-P_{mean})^2} \quad ...(18)$$

where O_i is the observed power generation, P_i is the predicted power generation, $\bar{O}$ and $\bar{P}$ are their means respectively, and N is the number of data points. R^2 provides a good measure of performance if the model to be fitted is linear but may not be an ideal goodness-of-fit statistic for nonlinear type of models. Another point of concern when using R^2 is that it tends to rise with the number of independent variables being considered, even if they may be redundant for the purpose.

The root mean square error (RMSE) is a commonly used metric for comparing model predictions with observed results. RMSE, which is expressed in the same units as the forecast variable, is of particular significance because of its property to attach a higher weightage to large errors which are unacceptable.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i-O_i)^2}{N}} \quad ...(19)$$

RMSE = 0 indicates a perfect fit and it increases with the variance of errors, the difference between predicted and observed values.

The Mean Absolute Error (MAE) is simply the average of absolute deviations of predicted values from the corresponding observed figures. Since it is the mathematical mean, the weightage attached to each deviation is the same, which is in contrast to the RMSE. Therefore, it is not susceptible to outliers in the data. MAE is determined using the following equation.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i| \quad ...(20)$$

As the above three metrics are capable of highlighting different characteristics, all of them were calculated to assess the performance of each statistical-, machine learning-, and hybrid-model employed in this study.

RESULTS AND DISCUSSION

Performance of the models

In the MLR- and PR-based models, the wind energy was expressed in terms of the wind speed (WS) and the ambient temperature (T) as follows.

$$E_{\mathrm{MLR}} = -5333.5 + 281\,\mathrm{WS} + 125.5\,\mathrm{T} \quad ...(21)$$

$$E_{\mathrm{PR}} = 2.84 \times 10^{-7}\ \mathrm{WS}^{1.71}\ \mathrm{T}^{5.16} \quad ...(22)$$

In the SARIMA model, the wind energy at time *t,*

$$E_{t(\mathrm{SARIMA})} = -41.45 + E_{t-12} + Z_t + 0.7602Z_{t-12} \quad ...(23)$$

where, Z_t denotes the forecast error at time t.

The variation of the predicted wind energy against the observed values corresponding to each prediction model is illustrated in Figure 8.

The performance of all the wind energy prediction models was compared in terms of the coefficient of determination, RMSE, and the MAE. As per the evaluation criteria, all the models are highly accurate. It can be clearly noticed that the two machine learning models of SVR and ANN, Figures 8(e) and 8(f), have produced the best linearity between the observed and predicted wind energies, as verified by the performance indicators in Table 2. This is closely followed by the performance of the three statistical techniques of MLR, PR and VEC whose scatter plots also show good agreement between observed and predicted values [Figures 8(a) and 8(b)]. All those five models have the common feature that weather indices of wind speed and ambient temperature were used as the input variables in modelling. The time series models, which are purely based on past wind energy data, also exhibit less error in terms of MAE and RMSE thus providing viable alternatives to the other prediction models. The single SVM model in which only the past energy data were used shows the least performance with the largest error which is also acceptable since it is a small portion of the energy.

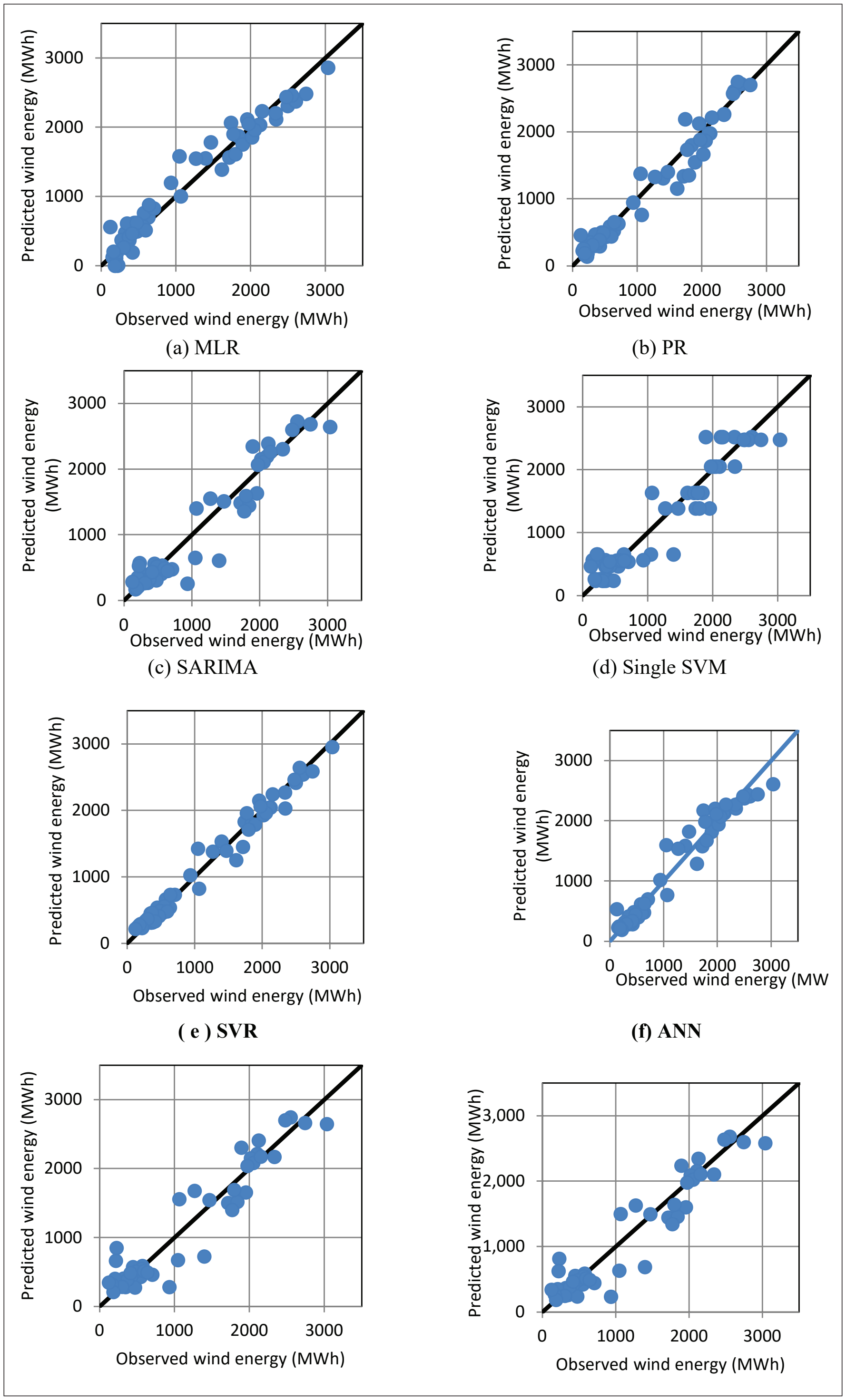

Figure 8: Scatter plots of predicted wind energy against the observed values with the line of best fit corresponding to each prediction model (a) MLR; (b) PR; (c) SARIMA; (d) Single SVM; (e) SVR; (f) ANN; (g) hybrid SARIMA-ANN; (h) Hybrid SARIMA-SVM

Table 2: Performance of the prediction models

	MAE	RMSE	R^2
SVR	77	133	0.98
ANN	123	173	0.96
MLR	142	177	0.95
PR	125	184	0.93
VEC	176	223	0.95
SARIMA	180	250	0.92
Single SVM	194	267	0.90
Hybrid SARIMA-SVM	190	264	0.91
Hybrid SARIMA-ANN	193	261	0.91

Subtle differences among model performances can be fathomed as the RMSE penalizes large prediction errors compared to MAE which is not sensitive to outliers. In this sense, SARIMA, SVM, and the two hybrid models indicate larger errors with their RMSE being 250 or more relative to the other models. More specifically within the former set of models, Figures 8(g) and 8(h) illustrate how points on graphs in hybrid models deviate more from the line of best fit relative to other models, implying some noteworthy errors in the hybrid models too. Further, of the two hybrid models, SARIMA-ANN performs slightly better than the SARIMA-SVM in terms of RMSE as their degree of collinearity denoted by R^2 is the same.

However, one may arrive at the opposite conclusion between the two hybrid models if judged by the MAE, as it is smaller in SARIMA-SVM. This is because MAE does not indicate the relative size of the error, which makes it difficult to distinguish large errors from small errors, since it is based on the ordinary mean value. As the model error magnitudes of SARIMA-SVM are unevenly distributed in the range from 0 to 790 MWh, it leads to an increase in the variance of the frequency distribution of error magnitudes. Therefore, the conclusion based on RMSE will be more appropriate than the contradictory one based on MAE. A similar clarification could be made to explain the different conclusions one may reach between the performances of the two models MLR and PR based on their corresponding metric values in Table 2. As the error magnitudes in both MLR and PR range from 2.5 to 785 MWh, more attention should be paid to the metric which attaches more weight to large errors, *i.e.*, the RMSE. Further, since PR is not a linear model, R^2 may not precisely project its performance as it is not the best goodness-of-fit statistic for non-linear type of regression models.

Looking at PR and VEC performances in terms of MAE and RMSE, PR is obviously the better model, but one may conclude the other way round if only the R^2 is used as the performance indicator. But R^2 is not the appropriate metric for model comparison here because in the VEC model, not only the wind speed and temperature, but also the past values of wind energy were used for modelling, and therefore the higher number of independent variables considered in VEC could well be the reason for yielding a higher R^2.

Considering the above analysis and if there is a contradiction for a choice based on MAE and RMSE, choosing the models indicated by RMSE is more appropriate in this research as the errors in all the models are not distributed evenly. Despite the slight differences within the model sub groups clarified above, machine learning models have outperformed other models for wind energy predictions at the Nala Danavi wind farm, which may be generalized to other coastal wind farms in the island exhibiting an annual pattern of weather and energy data similar to those in this study, and further, regression approaches like MLR, PR, and VEC, which use environmental factors, also lead to good prediction results.

Discussion

According to the performance of the models, SVR (R^2 = 0.98) and ANN (R^2 = 0.96) with very low error outperform the statistical and hybrid models. The superiority of SVR and ANN is achieved due to the non-linearity facilitated by machine learning and the use of weather indices as inputs in modelling. The statistical models of MLR, PR, and VEC, which used the same modelling approach, are also highly accurate with $R^2 \geq 0.93$. In contrast, the models which used only the past energy data in modelling yielded $0.92 \geq R^2 \geq 0.90$. These results indicate that developing prediction models based on the weather indices is more suitable. However, prediction of wind energy requires weather data to be projected, which needs additional work. Therefore, when the future weather data are unavailable, SARIMA, Single SVM, Hybrid SARIMA-SVM or Hybrid SARIMA-ANN models can be used.

Further to the detailed statistical analysis on the delicate differences among the models carried in the previous section, the overall performance of the present work was also compared with some similar studies reported in recent literature which are summarized in Table 3.

Table 3: Comparison of prediction models applied on wind farms

Reference	Country	Modelling techniques	Performance of the best model
Chen and Folly (2018)	South Africa	ANN, ANFIS, ARMA	RMSE = 5.6% MAE = 3.8%
Sfetsos (2000)	Greece	ARIMA, Linear Neural Network, back propagation NN, Levenberg Marquardt NN, RBF, Elman Recurrent Neural Network, ANFIS, (NLN), hybrid Logic Rules+ NLN	RMS error 4.89% better than the persistent method
Yona *et al.* (2013)		Auto Regressive, Kalman Filter, Neural Network	MAE = 68.732
Dupré *et al.* (2020)	France	ANN, ARMA	Δ_{RMSE} = 0.1%
Tena García *et al.* (2019)	Mexico	ANN, ARMA	MAE = 0.060 MSE = 0.079
Biswas *et al.* (2021)	US	ARIMA, Random Forest (RF), Bagging Classification & Regression Trees (BCART), ARIMA-RF, ARIMA-BCART	Boosted the prediction accuracy by 32%
Ekanayake *et al.* (2021)	Sri Lanka	ANN, MLR, PR	R = 0.97 RMSE = 109 Bias = -0.0003 Nash = 0.98
Zafirakis *et al.* (2019)	Greece	ANN, SVR	R^2 = 99.52% index of agreement = 99.71%

As understood from the comprehensive literature survey, only a few research studies have explored the efficacy of several wind energy prediction models falling under different categories. This research is unique in the context that nine types of statistical, machine learning, and hybrid techniques are applied with a comparison among them, and the best modelling options were proposed for wind energy prediction at a major wind farm in Sri Lanka. More importantly, the performances of the best models developed in this research are comparable to those of the models proposed by researchers on wind farms located in other countries.

CONCLUSION

As the prediction of wind power generation has been identified as extremely useful in managing power grids, this study focused on presenting prediction models for a key wind farm installation in Sri Lanka in the context of promoting clean energy sources. A total of nine models developed by applying statistical-, machine learning-, and hybrid-techniques were used with five years of data on wind energy, wind speed, and ambient temperature since commissioning of the wind farm. In general, all the models showed very high accuracy measured in terms of the MAE, RMSE, and R^2. Nevertheless, the difference in model performance of the three categories (statistical-, machine learning-, and hybrid-techniques) shall provide flexibility in choosing an appropriate model for a given situation depending on the availability of weather data. Due to the capacity in dealing with nonlinearity of data, the SVR model, which uses machine learning techniques, turned out to be the most accurate for predicting the wind energy when the weather data are available as inputs.

As indicated by the next higher values of R^2 and lower error measures of MAE and RMSE, the statistical models of MLR, PR, and VEC could also be used as either substitutes or alternatives for predicting wind energy within the same context of data availability. However, all the above models require weather data to be forecast first, based on which the wind energy is predicted. The other machine learning model of SVM, the time series model of SARIMA, and the two hybrid models of SARIMA-SVM and SARIMA-ANN can also be used for wind energy prediction in the event that future weather data are not available for some reason. The better accuracy of the single models over the hybrid models used in this study reinforced the findings of some previous studies that hybrid models do not always produce the best predictions for wind speed and power generation over the single forecasting models. Therefore, the results of this study also challenge the plausible opinion reported in some literature that hybrid modelling is superior to using individual modelling techniques. Rather, the findings of this study provide enough evidence to conclude that modelling techniques making the optimum use of past weather data would be the most appropriate for wind energy prediction at the Nala Danavi wind farm in Sri Lanka.

Conflicts of interest statement

The authors declare that they have no conflicts of interest.

Acknowledgement

Authors are thankful to LTL Holdings (Pvt) Ltd. and Ceylex Renewables (Pvt) Ltd. for providing data.

REFERENCES

Agarwal P., Shukla P. & Sahay K.B. (2018). A review on different methods of wind power forecasting. *Proceedings of the 2018 International Electrical Engineering Congress*, 7–9 March, Krabi, Thailand, pp. 1–4.
DOI: https://doi.org/10.1109/IEECON.2018.8712262

Bhatt G.A. & Gandhi P.R. (2019). Statistical and ANN based prediction of wind power with uncertainty. *Proceedings of the 3rd International Conference on Trends in Electronics and Informatics,* 23–25 April, Tirunelveli, India, pp. 622–627.
DOI: https://doi.org/10.1109/ICOEI.2019.8862551

Biswas A.K., Ahmed S.I., Bankefa T., Ranganathan P. & Salehfar H. (2021). Performance analysis of short and mid-term wind power prediction using ARIMA and hybrid models. *2021 IEEE Power and Energy Conference at Illinois (PECI)*, 01–02 April, Urbana, Illinois, USA, pp. 1–7.
DOI: https://doi.org/10.1109/PECI51586.2021.9435209

Box G.E.P. & Jenkins G. (1976). *Time Series Analysis: Forecasting and Control*, pp. 575. Holden-Day, Inc., San Francisco, CA, USA.

Chandra D.R., Kumari M.S. & Sydulu M. (2013). A detailed literature review on wind forecasting. *Proceedings of 2013 International Conference on Power, Energy and Control, (ICPEC)*, 6–8 February, Dindigul, India, pp. 630–634.
DOI: https://doi.org/10.1109/ICPEC.2013.6527734

Chang W.Y. (2014). A literature review of wind forecasting methods. *Journal of Power and Energy Engineering* **2**(04): 161–168.
DOI: https://doi.org/10.4236/jpee.2014.24023

Chen Q. & Folly K.A. (2018). Wind power forecasting. *IFAC-Papers OnLine* **51**(28): 414–419.
DOI: https://doi.org/10.1016/j.ifacol.2018.11.738

Dupré A., Drobinski P., Alonzo B., Badosa J., Briard C. & Plougonven R. (2020). Sub-hourly forecasting of wind speed and wind energy. *Renewable Energy* **145**: 2373–2379.
DOI: https://doi.org/10.1016/j.renene.2019.07.161

Ekanayake P., Peiris A.T., Jayasinghe J.M. & Rathnayake U. (2021). Development of wind power prediction models for Pawan Danavi wind farm in Sri Lanka. *Mathematical Problems in Engineering* **2021**: 4893713.
DOI: https://doi.org/10.1155/2021/4893713

Foley A.M., Leahy P.G. & McKeogh E.J. (2010). Wind power forecasting and prediction methods. *Proceedings of the 9th International Conference on Environment and Electrical Engineering,* 16–19 May, Prague, Czech Republic, pp. 61–64.
DOI: https://doi.org/10.1109/EEEIC.2010.5490016

Global Wind Energy Council (2020). *Global Wind Report 2019*. Global Wind Energy Council, Brussels, Belgium.

Haddad M., Nicod J., Mainassara Y.B., Rabehasaina L., Al Masry Z.A. & Péra M. (2019). Wind and solar forecasting for renewable energy system using SARIMA-based model. *International conference on Time Series and Forecasting*, 25–27 September, Gran Canaria, Spain, pp. 599–610.

Hanifi S., Liu X., Lin Z. & Lotfian S. (2020). A critical review of wind power forecasting methods - past, present and future. *Energies* **13**(15): 3764.
DOI: https://doi.org/10.3390/en13153764

Hansen J.V. & Nelson R.D. (2003). Time-series analysis with neural networks and ARIMA-neural network hybrids. *Journal of Experimental and Theoretical Artificial Intelligence* **15**(03): 315–330.
DOI: https://doi.org/10.1080/0952813031000116488

Huang Z. & Chalabi Z.S. (1995). Use of time-series analysis to model and forecast wind speed. *Journal of Wind Engineering and Industrial Aerodynamics* **56**(02-03): 311–322.
DOI: https://doi.org/10.1016/0167-6105(94)00093-S

Lehr J.H. & Keeley J. (2016). *Alternative Energy and Shale Gas Encyclopedia*, 1st edition, John Wiley & Sons, USA.
DOI: https://doi.org/10.1002/9781119066354

Li L.L., Zhao X., Tseng M.L. & Tan R.R. (2020). Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. *Journal of Cleaner*

Production **242**(05): 118447.
DOI: https://doi.org/10.1016/j.jclepro.2019.118447

Liu H., Tian H., Chen C. & Li Y. (2010). A hybrid statistical method to predict wind speed and wind power. *Renewable Energy* **35**(08): 1857–1861.
DOI: https://doi.org/10.1016/j.renene.2009.12.011

Lopes V.V., Scholz T., Estanqueiro A. & Novais A. Q. (2012, May). On the use of Markov chain models for the analysis of wind power time-series. *Proceedings of the 11th International Conference on Environment and Electrical Engineering,* 18–25 May. Venice, Italy, pp. 770–775.
DOI: https://doi.org/10.1109/EEEIC.2012.6221479

Mapuwei T.W., Bodhlyera O. & Mwambi H. (2020). Univariate time series analysis of short-term forecasting horizons using artificial neural networks: the case of public ambulance emergency preparedness. *Journal of Applied Mathematics* **2020**(02): 1–11.
DOI: https://doi.org/10.1155/2020/2408698

Milligan M., Schwartz M.N. & Wan Y.H. (2003). Statistical wind power forecasting for US wind farms. *Proceedings of the 17th Conference on Probability and Statistics in the Atmospheric Sciences*, 11–15 January, American Meteorological Society Annual Meeting Seattle, Washington, USA.

Santhosh M., Venkaiah C. & Vinod Kumar D.M.V. (2020). Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review. *Engineering Reports* **2**(06): 1–20.
DOI: https://doi.org/10.1002/eng2.12178

Sfetsos A. (2000). A comparison of various forecasting techniques applied to mean hourly wind speed time series. *Renewable Energy* **21**(01): 23–35.
DOI: 10.1016/S0960-1481(99)00125-1

Shi J., Guo J. & Zheng S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews* **16**(05): 3471–3480.
DOI: https://doi.org/10.1016/j.rser.2012.02.044

Sideratos G. & Hatziargyriou N.D. (2007). An advanced statistical method for wind power forecasting. *IEEE Transactions on Power Systems* **22**(01): 258–265.
DOI: https://doi.org/10.1109/TPWRS.2006.889078

Tena García J.L., Cadenas Calderón, E., González Ávalos, G., Rangel Heras E. & Mbikayi Tshikala A. (2019) Forecast of daily output energy of wind turbine using SARIMA and nonlinear autoregressive models. *Advances in Mechanical Engineering* **11**(2): 1–15.
DOI: https://doi.org/10.1177/1687814018813464

Tseng F.M., Yu H.C. & Tzeng G.H. (2002). Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change* **69**(01): 71–87.
DOI: https://doi.org/10.1016/S0040-1625(00)00113-X

Voyant C., Muselli M., Paoli C. & Nivet M.L. (2012). Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. *Energy* **39**(01): 341-355.
DOI: https://doi.org/10.1016/j.energy.2012.01.006

Wang X., Guo P. & Huang X. (2011). Energy procedia a review of wind power forecasting models. *Energy Procedia* **12**: 770–778.
DOI: https://doi.org/10.1016/j.egypro.2011.10.103

Yona A., Senjyu T., Toshihisa F. & Kim C.H. (2013). Very short-term generating power forecasting for wind power generators based on time series analysis. *Smart Grid and Renewable Energy* **4**(02): 181-186.
DOI: https://doi.org/10.4236/sgre.2013.42022

Zafirakis D., Tzanes G. & Kaldellis J.K. (2019) Forecasting of wind power generation with the use of artificial neural networks and support vector regression models. *Energy Procedia* **159**: 509–514.

Zhao Y., Ye L., Li Z., Song X., Lang Y. & Su J. (2016). A novel bidirectional mechanism based on time series model for wind power forecasting. *Applied Energy* **177**(C): 793–803.
DOI: https://doi.org/10.1016/j.apenergy.2016.03.096