21st Australasian Fluid Mechanics Conference
Adelaide, Australia
10-13 December 2018

Wind Speed Forecasting using Regression, Time Series and Neural Network Models: a Case Study of Kiribati

A. Arzu[1], S. S. Kutty[2], M.R. Ahmed[2] and M.G.M. Khan[1]

[1]School of Computing, Information and Mathematical Sciences
The University of the South Pacific, Suva, Fiji

[2]School of Engineering and Physics
The University of the South Pacific, Suva, Fiji

Abstract

There is an increase in demand for renewable sources of energy due to apprehensions about climate change, increase in the energy demand and unpredictability of the prices and supply of fossil fuels. Wind energy is one of the world's fastest growing sources of energy. As a result of the stochastic behavior of wind, the demand for accurate wind forecasting has become imperative to reduce the risk of uncertainty.

In this paper, the wind speed data are modelled and forecasted using three forecasting techniques: Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN). To test these models for wind speed forecasting, daily wind speed, pressure, relative humidity and temperature data for the period of September 2012 to September 2013 for Abaiang in Kiribati were used in this work. The performance of the models was evaluated using four measures: root mean square error, mean absolute error, mean absolute percentage error and coefficient of determination (R^2). The optimum model was also compared to a benchmark technique, persistence method. The empirical results reveal that the proposed model using Artificial Neural Network is more efficient and accurate in forecasting wind speed in comparison to the regression and time series models.

Introduction

The energy sector encompasses approximately two-thirds of global total greenhouse gas emissions which contribute to the climate change phenomena that has become a global concern [8]. While climate change causation is global, some regions are being affected more than others. Pacific islands countries (PIC), particularly those in warmer regions, are the most susceptible to its effects [10]. The contribution of the PICs is below 0.03% of current global greenhouse gas emissions according to UN Permanent Forum on Indigenous Issues [10], yet they are among the first to be affected and it is projected that their populations will be among the first that will need to adapt to climate change or be subjected to relocation and abandonment of their traditional homelands. Island states such as Kiribati that are only slightly above sea level are the immediate victims of this phenomenon.

The Republic of Kiribati (pronounced Kiribas) formerly known as the Gilbert Islands, is located in the center of the Pacific. The nation consists of 33 small islands divided into three distinct archipelagos: Gilbert Islands, Phoenix Islands and Line Islands. All 33 islands are low-lying atolls or reef-top islands except Banaba, which is a raised atoll west of the Gilberts [11].

Renewable energy resources are being increasingly capitalized on to provide greenhouse gas emission-free sources of electricity in order to lessen the effect on climate change. Wind energy has become the world's fastest growing renewable energy source of electricity generation because as described by Li and Shi [6] wind energy is "socially beneficial, economically competitive, and environmentally friendly" (p. 2313). Compared with fossil fuels, wind energy has its unique characteristics such as low energy density, randomness, instability and volatility, rapid changes in wind direction and magnitude, and is easily influenced by the geographical conditions and the surrounding environment [12]. The abundance of wind energy does not compensate for its stochastic behaviour. Forecasting is a necessity to decrease the risk of ambiguity and allowing better incorporation of wind energy into power systems [2].

Considerable research efforts have been directed at developing superior forecasting methods. The most common methods for forecasting wind speed include: Persistence Approach, Physical Approach, Statistical Approach and Hybrid Approach. Persistence Approach assumes that the wind speed at a certain time in the future will be the same when the forecast is made and is normally utilized as a benchmark for comparing other short-term wind speed forecast tools [13]. Physical Approaches like Numeric Weather Prediction (NWP) use parameterizations stemming from a comprehensive description of the physicality of the atmosphere such as terrain, obstacle, pressure, and temperature to estimate the future wind speed [3,12]. Statistical Approaches such as Time-Series models, Regression models and Artificial Neural Networks are based on training with data measured and utilizes errors to adjust the parameters of the model [2]. Hybrid Approach generally combines different method e.g. combining physical and statistical approaches while maintaining the strength of each method to enhance the performance of the forecasting model [5,13].

Selecting appropriate input variables is essential in building an effective forecasting model. Different variables are required for different models. Physical models use physical considerations to forecast wind speed, therefore the input variables will be the physical or meteorological information while Statistical models use historical wind speed data and NWP output as input. Understanding the importance and relevance of the parameters that affect wind speed is important when choosing inputs. Wind speed, relative humidity, power generation hours, mean temperature, wind gust, wind direction and barometric pressure have been used as inputs. Once a strong correlation between wind speed and other variables is established, these variables can be used along with wind speed as inputs in order to help in the prediction of wind speed. Statistical method is cost effective and the technique is generally selected for short term forecast. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN) are the three statistical methods modelled in this research.

Statistical Models

Multiple Linear Regression (MLR)

The multiple linear regression (MLR) model is defined by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (1)$$

where *y* is the dependent variable wind speed, $x_i; (i = 1, 2, ..., k)$ are the predictor variables, and $\beta_0, \beta_1, ..., \beta_k$ are regression coefficients, *k* is the number of predictor variables and ε is the vector of residuals. The model assumes that the residuals are normally distributed with mean zero and variance is constant.

Once the model has been established, the assumptions are tested to determine the robustness of the model. Linearity was assessed graphically using Residuals versus Predicted plot, independence was assessed using Durbin-Waston test, normality was assessed graphically through the Q-Q Plot and Shapiro-Wilk test and homoscedasticity was determined using Heteroskedacity Test and graphical plots of residuals versus each independent variable.

Autoregressive Integrated Moving Average (ARIMA)

The Box and Jenkins iterative procedure for modeling a time series was used as postulated by [1,7]. This iterative modeling approach encompasses three phases:

i) Identification, in which the characteristics and statistics of a time series are examined. ARIMA models require the input data to have a constant mean, variance, and autocorrelation through time. The stationarity of the input data series is determined via the autocorrelation function (ACF) and Partial Autocorrelation (PACF) tests. The unit root test can also be used to determine stationarity.

ii) Estimation, in which we estimate the parameters of potential model(s) using the data at hand.

iii) Diagnostic checking, in which we examine the estimated model(s), and residuals of the fitted model(s), to see if the model(s) make sense and are in agreement with our assumptions.

The general non-seasonal model is also known as ARIMA (*p, d, q*), where *p* is the order of the autoregressive part of the model, *d* is the order of differencing done to the data to make it stationary and *q* is the order of the moving average part of the model.

The ARIMA model is defined by:

$$y_t = \sum_{i=1}^{P} \varphi_i y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + \varepsilon_t \qquad (2)$$

where φ_i is the *i*th autoregressive parameter, θ_j is the *j*th moving average parameter and ε_t is the error term at time t.

Artificial neural network (ANN)

The ANN model designed is a multi-layered perception (MLP). The proposed model considers the most widely used neural network, known as the back propagation network. The necessary components needed to establish a neural network is outlined by Cadenas and Rivera [1] as follows:

i) Its architecture (the number of layers and units in the network and connections among them). An ANN is typically composed of layers of nodes. Most applications need networks that contain three or more layers – input, hidden, and output. In the MLP, all input nodes are in one input layer, all output nodes are in one output layer and the hidden nodes are distributed into one or more hidden layers**.**

ii) The activation function (that describes as each unit combines its inputs to obtain the desired outputs). The activation functions below are used in this research to determine which would produce optimal results:

a) Sigmoid (logistics) function

$$f(x) = \frac{1}{1 + e^{-x}}$$

b) Cosine function

$$f(x) = \cos(x)$$

c) Sine function

$$f(x) = \sin(x)$$

d) Tangent hyperbolic function

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

iii) The cost function (a measurement of the accuracy of the prediction). Typically, Sum Squared Error, $SSE = \sum_{t=1}^{n} (E_t)^2$, and Mean Squared Error, $MSE = \sum_{t=1}^{n} (E_t)^2 / n$ are used since they are defined in terms of error, $E_t = \hat{y}_t - y_t$, in the output and the hidden layers, where $\hat{y}_t$ is the final output at an output layer, y_t is the actual value of the output node for the t th observation $(t = 1, 2, ..., n)$. This is important because optimality of model is defined by least error. In this research, the SSE is utilized.

iv) The training algorithm to find the values of the parameters that diminish the cost function. The application algorithm for backpropagation, outlined below, is by Sivanandam & Deepa [9] as:

Step 1: Initialize weights (from training algorithm)
Step 2: For each input vector do steps 3-5.
Step 3: For *i* = 1,…*n*; set activation of input unit, x_i;

Step 4: For *j* = 1,…*p*; $Z_{-inj} = v_{oj} + \sum_{i=1}^{n} x_i v_{ij}$ (3)

Step 5: For *k*=1,…*m*; $y_{-ink} = w_{ok} + \sum_{j=1}^{p} z_j w_{jk}$ (4)

$$y_k = f(y_{-ink}) \qquad (5)$$

where *x* is the input training vector, z_j is the hidden unit *j*, v_{oj} is the bias on hidden unit *j*, w_{ok} is the bias on output unit *k* and y_k is the output unit *k*.

Model Evaluation

The development of every model requires comparisons with other models to distinguish accuracy and superiority. While the accuracy of forecasting is important, the measures for the evaluation of forecasting models are also important. A number of performance measures have been used to evaluate the forecast accuracy, but there is still a debate on which is best or which is recognized as the universal standard. According to Hyndman and Koehler [4] many of the recommended measures were found to be inadequate, and many of them degenerate in commonly occurring situations suggesting that Mean Square Error and Root Mean Square Error are largely used because of their theoretical relevance in statistical modelling; however authors have dissuaded the use because they are more sensitive to outliers. To quantitatively determine the optimal model, three forecast error measures are employed for model comparison and evaluation: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Error measures can be defined as:

i) MAPE = $\% \frac{\frac{\sum |e(t)|}{x(t)}}{n}, t = 1, 2, ..., n$ (6)

$$\text{ii) RMSE} = \sqrt{\frac{\sum |e(t)|^2}{n}}, t = 1, 2, ..., n \quad (7)$$

$$\text{iii) MAE} = \frac{\sum |e(t)|}{n}, t = 1, 2, ..., n \quad (8)$$

where: $x(t)$: actual data at time t, $f(t)$: the estimate of forecast of time t and $e(t)$: predicted error at time t, $e(t) = x(t) - f(t)$.

The fitness of data will be measured using the Coefficient of determination:

$$R^2 = \quad 1 - SSE/SST \quad (9)$$

where SSE(Sum of Squares due to error and SST is the total sum squares.

Data measurement, Inputs and Parameter

The data used herein were provided by the University of the South Pacific, Engineering Department. Measurements were performed at a height of 34 m above ground level over a period of approximately 1 year; from September 2012 to September 2013 on the island of Abaiang in Kiribati. The data sampling interval was 10 minutes. Table 1 shows the variables measured and used in this study and Figure 1 illustrates the time series plots of the daily wind speeds.

Variables	Units
Wind speed	m/s
Direction	Deg
Temperature	Deg C
Humidity	%RH
Pressure	mBar

Table 1. List of Parameters.

Variable selection is dependent on recognizing relationships within the data that are suitable predictors of the model output. Pearson's Correlation defined by:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (10)$$

was used to identify useful explanatory variables, that is, variables showing significant correlation with wind speed. Results presented in Table 2 show that pressure, direction and humidity are statistically significant predictor variables for the development of the wind speed forecasting models.

The dataset was further divided into three subsets: training, validation, and testing datasets. The data contains 365 observations for each variable of which 292 observations (80%) were used for training the models and the remaining 20% was divided in validating and testing sets with 37 and 36 observations respectively.

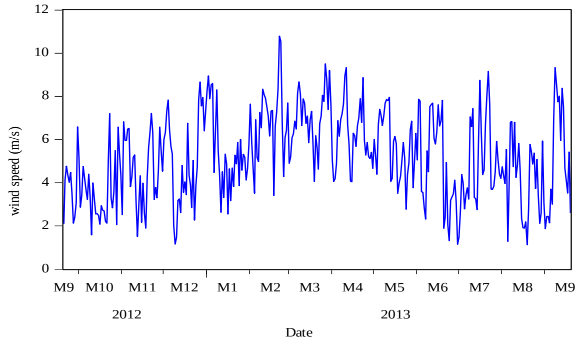

Figure 1. Time series plot of Abaiang's wind speed.

	Speed	Direction	Temperature	Humidity	Pressure
Speed	1	-0.452**	-0.020	0.178**	-0.357**
Direction	-0.452**	1	-0.141	-0.290**	0.036
Temperature	-0.020	-0.141	1	-0.041	0.000
Humidity	0.178**	-0.290**	-0.041	1	0.390**
Pressure	-0.357**	0.036	0.000	0.390**	1

**Correlation is significant at the 0.01 level (2-tailed).

Table 2. Correlation Analysis of Abaiang Input Data.

The Artificial Neural Network requires additional pre-processing in the form of normalization. The normalization measure is defined as:

$$X' = \{X'_i\} = 2 \times \left(\frac{X_i - \min X_i}{\max X_i - \min X_i} \right) - 1 \quad (11)$$

where $i = 1, 2, ..., n$ and $X' \subset [-1, 1]$, min X_i and max X_i are the minimum and maximum value of the input array and X_i denotes the real value of each vector.

Results and Discussion

The parameters of the best model for each forecasting method are shown in Table 3. Figures 2, 3 and 4 depict the actual, predicted and residual wind speed values of daily average for MLR, ARIMA and ANN models respectively. It can be seen that all the models capture a similar tendency of the actual data.

Model	Parameter
MLR	3 lags, direction, relative humidity, pressure
ARIMA	Autoregressive model of order p = 3, degree of differencing d = 0 and moving-average model of order q =3, that is ARIMA(3,0,3).
ANN	6 inputs variables (4 lags, direction, relative humidity), 2 hidden nodes and 1 output node with Sigmoid activation function.

Table 3. Parameters of MRL, ARIMA and ANN.

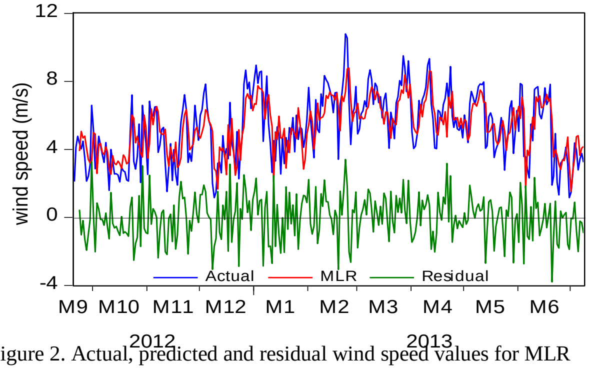

Figure 2. Actual, predicted and residual wind speed values for MLR training sample.

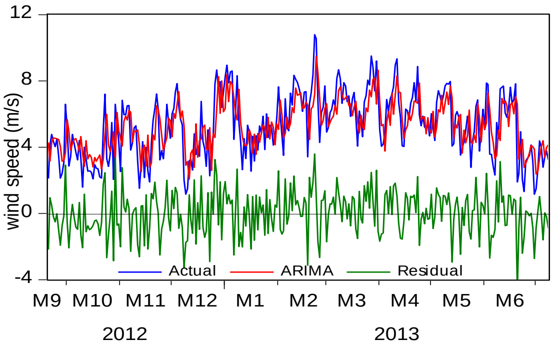

Figure 3. Actual, predicted and residual wind speed values for ARIMA training sample.

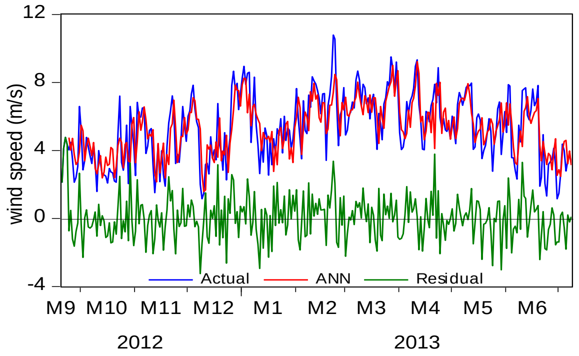

Figure 4. Actual, predicted and residual wind speed values for ANN training sample.

From the results, it can be seen that the models trained with varying parameter inputs result in varying degrees of accuracy. The training and validation sets were used to construct and identify the optimum model for each forecasting type. The validation set was particularly significant for the ANN model construction to avoid overfitting.

The models further undergo a comparative analysis to identify which forecasting technique gives superior forecasting of Abaiang by comparing with a benchmark technique, persistence method. The test dataset was used for this process. The use of the test set is crucial because it was not used in the model fitting which permits genuine forecasts [1]. Figure 5 compares the actual and 36 days predicted values obtained by using the MLR, ARIMA, ANN and Persistence models respectively. It can be seen that all the models capture a similar tendency of the actual data. Table 3 shows the statistical error measures: RMSE, MAE, MAPE and variation of data capture measure R^2 for the MLR, ARIMA, ANN and Persistence models. It can be observed that the error values obtained with the ANN model are considerably lower than other models and the R^2 is also the highest for ANN.

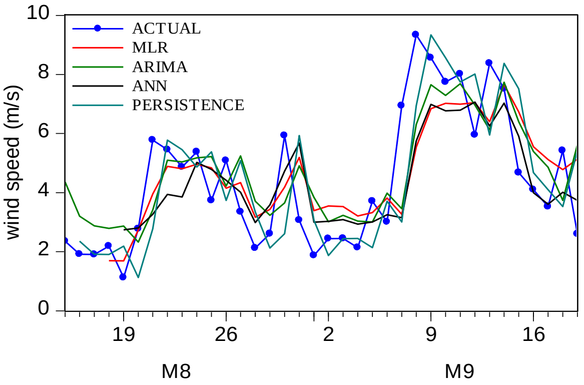

Figure 5. Actual, MLR, ARIMA, ANN and Persistence predicted test values.

Forecast Model	RMSE	MAE	MAPE	R^2
MLR 31	1.5439	1.2726	34.4138	0.4047
ARIMA (3, 0, 3)	1.5649	1.3162	39.5564	0.5117
ANN (6, 2, 1)	**1.4822**	**1.1863**	**29.7312**	**0.5505**
Persistence	1.6973	1.3451	34.5891	0.4281

Table 3. Statistical error measures for MLR, ARIMA, ANN and Persistence models of test set.

Conclusions

Wind energy is one of the world's fastest growing sources of energy. Improving forecasting accuracy is crucial in superior forecasting. Some researchers model hybrids, while others enhance and improve forecasts by including explanatory variables. The latter was used in this study to encompass wind speed wholly; therefore, it was important to include those external variables that affect wind speed. Lag wind speed, direction and humidity were identified as relevant variables for forecasting. Three forecasting techniques, namely Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) were employed to forecast the wind for a period of 36 days and compared with the actual data. The results showed that the three models reasonably forecasted Abaiang's wind speed in comparison to Persistence model (benchmark) with ANN predicting with a higher degree of accuracy. With this technique, the R^2 value was the highest and the error parameters RMSE, MAE and MAPE were the lowest.

References

[1] Cadenas, E., & Rivera, W. (2007). Wind speed forecasting in the south coast of Oaxaca, Mexico. *Renewable Energy, 32*(12), 2116-2128.

[2] Foley, A. M., Leahy, P. G., Marvuglia, A., & McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. *Renewable Energy, 37*(1), 1-8.

[3] Hu, J., Wang, J., & Zeng, G. (2013). A hybrid forecasting approach applied to wind speed time series. *Renewable Energy, 60*, 185-194.

[4] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting, 22*(4), 679-688.

[5] Jung, J., & Broadwater, R. P. (2014). Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews, 31*, 762-777.

[6] Li, G., & Shi, J. (2010). On comparing three artificial neural networks for wind speed forecasting. *Applied Energy, 87*(7), 2313-2320.

[7] Liu, L.-M., Hudak, G. B., Box, G. E., Muller, M. E., & Tiao, G. C. (1992). *Forecasting and time series analysis using the SCA statistical system* (Vol. 1): Scientific Computing Associates DeKalb, IL.

[8] Pryor, S. C., & Barthelmie, R. J. (2011). Assessing climate change impacts on the near-term stability of the wind energy resource over the United States. *Proceedings of the National Academy of Sciences of the United States of America, 108*(20), 8167-8171

[9] Sivanandam, S., & Deepa, S. (2006). *Introduction to neural networks using Matlab 6.0*: Tata McGraw-Hill.

[10] United Nations Forum on Indigenous Issues, 14 (2015). Together We Achieve. Retrieved from http://www.un.org/esa/socdev/unpfii/documents/2015/media/pacific.pdf

[11] Woodroffe, C. D. (2008). Reef-island topography and the vulnerability of atolls to sea-level rise. *Global and Planetary Change, 62*(1-2), 77-96.

[12] Wu, B., Song, M., Chen, K., He, Z., & Zhang, X. (2014). Wind power prediction system for wind farm based on auto regressive statistical model and physical model. *Journal of Renewable and Sustainable Energy, 6*(1), 013101.

[13] Zhang, W., Wang, J., Wang, J., Zhao, Z., & Tian, M. (2013). Short-term wind speed forecasting based on a hybrid model. *Applied Soft Computing, 13*(7), 3225-3233.