22nd Australasian Fluid Mechanics Conference AFMC2020
Brisbane, Australia, 7-10 December 2020
https://doi.org/10.14264/ccee311

# Wind Speed Forecasting using Regression, Time Series and Neural Network Models: A Case Study of Suva

**A. Arzu[1], M.R. Ahmed[2] and M.G.M. Khan[3]**

[1]Faculty of Science and Technology, Department of Mathematics, Physics and Information Technology

University of Belize, Belize

[2]School of Engineering and Physics

The University of the South Pacific, Suva, Republic of Fiji

[3]School of Computing, Information and Mathematical Sciences

The University of the South Pacific, Suva, Republic of Fiji

## Abstract

There has been an increase in the capitalization of renewable energy resources to provide greenhouse gas emission-free sources of electricity in order to lessen the effect on climate change. This increase is due to trepidations about climate change, an 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. Though abundant, its abundance does not compensate for its stochastic behavior. Forecasting is therefore necessary to increase efficiency and reduce 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). Four variables: daily wind speed, pressure, relative humidity and temperature were used to develop the wind speed forecast from these models. The performance of the models was evaluated using four measures: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination ($R^2$). Results show that the superior model to forecast wind speed is the Multiple Linear Regression. The empirical results reveal that the proposed model using Multiple Linear Regression is more efficient and accurate in forecasting wind speed in comparison to time series models.

*Keywords*

Wind energy; Forecasting; Artificial neural network; Time series models; Multiple linear regression.

## Introduction

Energy production and consumption forms 67% of the total emissions. It is the dominating component in global greenhouse gas emissions; which contributes to the climate change phenomena [13]. Climate change has amplified existing risks and continues to create new risks for natural and human systems. Unfortunately, these risks are unevenly distributed and are generally greater for disadvantaged people and communities. Pacific islands countries (PIC), particularly those in warmer regions, are the most susceptible to the effects of climate change [16]. The contribution of the PICs is below 0.03% of current global greenhouse gas emissions according to UN Permanent Forum on Indigenous Issues [16]; yet they are among the first to be affected and their populations will be among the first that will need to adapt to the adverse effects of climate change. Some islands are already facing climate change impacts on communities, infrastructure, water supply, coastal and forest ecosystems, fisheries, agriculture, and human health.

The Republic of Fiji, one of the PICs, is determined to exploit renewable energy to reduce its carbon footprint. Fiji is an archipelago of about 332 islands, lies in the heart of the South Pacific and occupies 18,376 km² of land mass. Approximately one third (110) of the islands are inhabited [3]. Fiji is divided into two geographic regions, the western region which is dominated by the two largest islands Viti Levu and Vanua Levu; accounts for about 88% of the total land area and 90% of the total population. The eastern region is made up of small islands, including the Lau Group. Suva, the capital of Fiji, is located on the southeastern coast of Vitu Levu [3].

As the capitalization of wind speed energy increases due to its social, economic, competitive and environmental-friendly properties, other lesser attractive properties such as its randomness, instability, volatility and rapid change in direction cannot be forgotten or dismissed [9]. Its stochastic behaviour cannot be compensated by its abundance; therefore, forecasting is necessary to decrease the risk of ambiguity and allowing better incorporation into power systems [5].

The most common methods for forecasting wind speed includes: Persistence Approach which is often considered the benchmark, assumes that the wind speed at time '$t+\Delta t$' will be the same as it was at time 't'; that is, the future wind speed is identical to the current one [13]. Physical Approach such as Numeric Weather Prediction (NWP) uses parameterizations founded on a detailed physical description of the atmosphere such as terrain, obstacle, pressure, and temperature to estimate the future wind speed [6]. Statistical Approach aims to find relationships based on training with data measured and utilizes errors to adjust the parameters of the model. These include: Time-Series models, Regression models and Artificial Neural Networks [5]. Hybrid Approach combines different methods while maintaining the strength of each method to enhance the performance of the forecasting model [8].

It is evident that wind speed depends on external variables, therefore it is very appropriate to develop the wind speed forecasting models on these external variables. Identifying appropriate input variables is crucial when building an effective forecasting model. The input variables for each model varies; physical models use physical considerations like terrain, pressure etc., 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. Inputs that have been used are: wind speed, relative humidity, mean temperature, wind gust, wind direction and barometric pressure. According to Castellanos and James [4], 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 Models

Statistical approaches are easy to model, inexpensive, and provides well-timed forecast aiming particularly at short-term forecasting [13]. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN) are the three statistical methods modelled in this work.

#### Multiple Linear Regression (MLR)

The multiple linear regression (MLR) model is defined by

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

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 $\varepsilon$ is the vector of residuals. The model assumes that the residuals are normally distributed with a mean of zero and variance is constant. After the construction of the model, its robustness is determined by verifying the validity of assumptions. Linearity was evaluated using graphical plots of Residuals versus predicted plot, the Durbin-Waston test was used to test independence, normality was assessed using Shapiro-Wilk test and graphically through the Q-Q Plot and, homoscedasticity was determined graphically using residuals versus each independent variable plots and the Heteroskedacity Test.

#### Autoregressive Integrated Moving Average (ARIMA)

The Box and Jenkins iterative procedure for modeling a time series was used as postulated by [2]. This iterative modeling approach encompasses the phases:

i) Identification: 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: the parameters of potential model(s) using the data at hand are estimated.

iii) Diagnostic Testing: the estimated model(s) and residuals of the fitted model(s) are examined to identify any inadequacies and determine if the residues are white noise.

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:

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

where $\varphi_i$ is the $i$th autoregressive parameter, $\theta_j$ is the $j$th moving average parameter and $\varepsilon_t$ is the error term at time t.

#### Artificial neural network (ANN)

A Multi-layered perception (MLP), ANN model was designed. 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 are outlined by Cadenas and Rivera [2] 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 measure of the accuracy of the prediction): typically SSE (sum squared error) and MSE (mean squared error) are used since they are defined in terms of error. This is important because optimality of model is defined by the least error. The SSE is utilized in this research.

iv) The training algorithm to find the values of the parameters that diminish the cost function. The application algorithm for backpropagation, outlined below, as presented by Sivanandam and Deepa [14] as:

***Step 1***: Initialize weights (from training algorithm)
***Step 2***: For each input vector, carry out steps 3-5.
***Step 3***: For $I = 1, \ldots n$; set activation of input unit, $x_i$;

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

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

The output is given by $Y_k = f(y_{-ink})$ (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

Accuracy and superiority in forecasting are of paramount importance when developing forecasting models. This requires comparisons with other models. While the accuracy of forecasting is important, the measures for the evaluation of forecasting models are equally important. Many performance measures have been used to evaluate the forecast accuracy, there is however an ongoing debate on which is best or which is recognized as the universal standard. According to Hyndman and Koehler [7]. many of the recommended measures were found to be inadequate, and many of them degenerate in commonly occurring situations. To quantitatively determine the optimal model, three forecasting 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:

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

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

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

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

The fitness of data is measured using the Coefficient of determination:

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

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

## Data measurement, Inputs and Parameter

The data were acquired at the University of the South Pacific's campus. Measurements collected from October 2012 to October 2013 were taken at 34 m above ground level with a data sampling interval of 10 minutes. The data were averaged to convert to daily data. The data consisted of wind speed (m/s), pressure (mBar), direction (Deg), humidity (%RH), and temperature (Deg C). While each variable should consist of 365 observations; only 92.9% of the data was available. Some values were missing for each variable except temperature. Table 1 summarizes each variable measured.

| Variables | N | Missing | | Min | Max |
|---|---|---|---|---|---|
| | | Count | % | | |
| Speed | 341 | 24 | 6.58 | 0.82 | 18.97 |
| Direction | 341 | 24 | 6.58 | 26.00 | 323.52 |
| Temperature | 365 | 0 | 0.00 | 19.96 | 31.70 |
| Humidity | 364 | 1 | 0.27 | 55.00 | 97.00 |
| Pressure | 363 | 2 | 0.55 | 992.83 | 1019.90 |

Table 1. List of Parameters.

It can also be observed that there is a high wind speed maximum of 18.97 m/s recorded during tropical cyclone Evan that made landfall on Fiji on December 17, 2012. Such occurrence is not frequent in Fiji; thus, special precaution was taken with the data. Zhang et al. [13] stipulated that while ANN outperforms linear regression models when outliers are present in data, ANN can still be vulnerable to these outliers. To eliminate such vulnerability, the data collected on December 12 were discarded and classified as missing data. The missing data were then imputed using the Multivariate Imputation by Chained Equations (MICE) package in the R software as detailed by Buuren and Groothuis-Oudshoorn [1]. Figure 1 shows time-series plot of wind speed of the completed data.

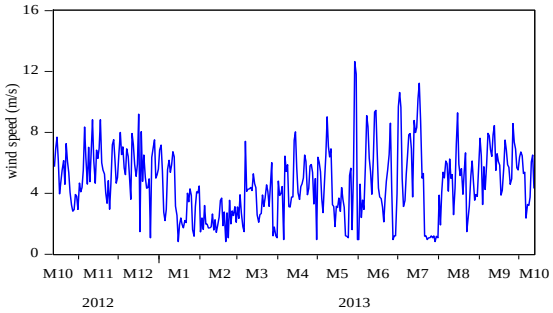


Figure 1. Time series plot of complete wind speed data for Suva.

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 ($r \neq 0$) with wind speed. Table 2 shows that pressure, direction and humidity are statistically significant predictor variables for the development of the wind speed forecasting models.

| | Speed | Direction | Temperature | Humidity | Pressure |
|---|---|---|---|---|---|
| **Speed** | 1 | -0.448** | -0.081 | -0.122* | 0.377** |
| **Direction** | -0.448** | 1 | -0.259** | 0.220** | -0.155** |
| **Temperature** | -0.081 | -0.259* | 1 | -0.202** | -0.383** |
| **Humidity** | -0.122* | 0.220** | -0.202** | 1 | 0.045 |
| **Pressure** | 0.377** | -0.155** | -0.383** | 0.045 | 1 |

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Table 2. Correlation Analysis of Suva Input Data.

Additional pre-processing; that is normalization, was required for the Artificial Neural Network model. 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 for MLR, ARIMA and ANN models respectively.

| Model | Parameter |
|---|---|
| MLR | 1 lag, direction, relative humidity |
| ARIMA | 3,0,3<br>Autoregressive parameter (p) = 1,<br>Differencing (d) = 0 and<br>Moving Average parameter (q) =3 |
| ANN | 7, 2, 1<br>Inputs: 7 (5 lags, direction, relative humidity)<br>Hidden nodes:2 and Output node: 1<br>Activation function: Sigmoid Function |

Table 3. Parameters of MRL, ARIMA and ANN

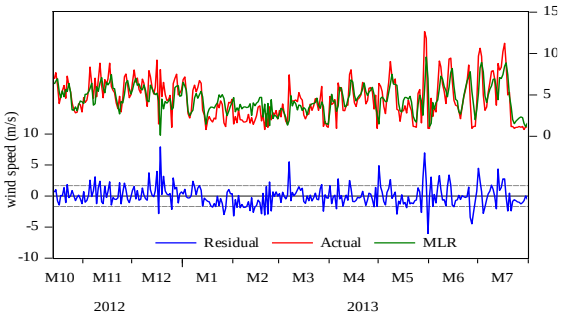


Figure 2. Actual, predicted and residual wind speed values for MLR training sample.

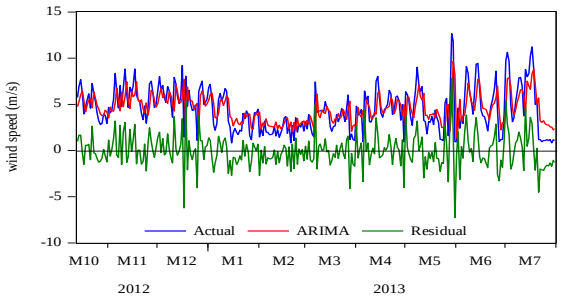


Figure 3. Actual, predicted and residual wind speed values for ARIMA training sample

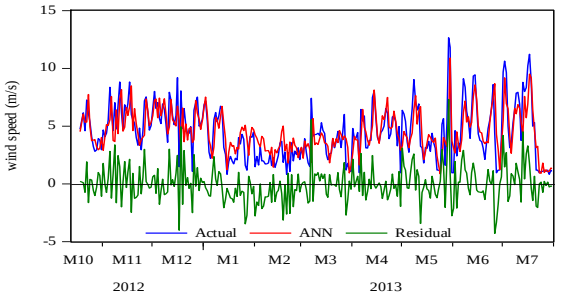


Figure 4. Actual, predicted and residual wind speed values for ANN training sample

Results show that the trend of Suva's wind speed was captured by each forecasting model. The optimum model for each forecasting model was identified by a thorough variation analysis of parameters. Each variation of parameter yielded different degrees of accuracy. Training and validation sets were used to construct the models. Validation sets were particularly important for the ANN modelling to prevent overfitting.

The test set consisting of 36 observations was then used to identify the optimum model. The use of the test set was crucial too because it was not used in the training; it allows genuine forecast [2]. Figure 5 compares the actual and predicted values of the test set using the MRL, ARIMA and ANN models. Table 4 shows the statistical error measures: RMSE, MAE, MAPE and variation of data capture measure $R^2$ for the MLR, ARIMA, ANN models. Noticeably, the errors for the MLR models were the lowest; however, the variation of data ($R^2$) captured is highest for ANN. It can be noted that ANN contains more input variables as compared to MLR and ARIMA, which may have influenced the higher $R^2$ value. Thus, the results reveal that the MLR model performs better than the ARIMA and ANN models and it can be considered a superior model for forecasting wind speed at the Suva location.

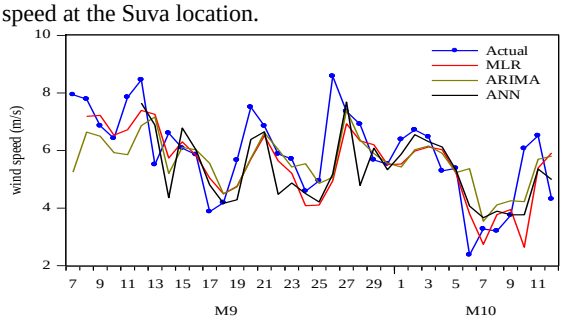


Figure 5. Actual, MLR, ARIMA, ANN predicted test values.

| Forecast Model | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|
| MLR 13 | **1.1697** | **0.8469** | **15.3565** | 0.3036 |
| ARIMA (1, 0, 3) | 1.2822 | 0.9401 | 17.6911 | 0.2575 |
| ANN (6, 2, 1) | 1.1828 | 0.8754 | 16.0476 | **0.3352** |

Table 4. Statistical error measures for MLR, ARIMA and ANN models of test set.

## Conclusion

Wind energy is becoming one of the world's fastest growing sources of energy. Improving forecasting accuracy is crucial in superior forecasting. One such way of enhancing and improving forecasting accuracy is by including explanatory variables. The use of explanatory variables in this study is an effort to study wind speed forecasting holistically; consequently, it was imperative to include those external variables that affect wind speed. Lag wind speed, direction and relative humidity were identified as relevant variables for forecasting. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) were the forecasting techniques employed to forecast the wind for a period of 36 days and compare with the actual data. The comparison results strongly indicate that the model MLR is an efficient tool for forecasting Suva's wind speed with the highest degree of accuracy.

## References


[1] Buuren, S., and Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3), 1-68.

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

[3] Clark, G., and Anderson, A. (2009). *TA31: The Early Prehistory of Fiji*. ANU Press.

[4] Castellanos, F., and James, N. (2009). *Average hourly wind speed forecasting with ANFIS.* Paper presented at the 11th Americas conference on Wind Engineering.

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

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

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

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

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

[10] Pryor, S. C., and 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

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

[12] 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

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