ARTIFICIAL NEURAL NETWORK AND KALMAN FILTER APPROACHES
BASED ON ARIMA FOR DAILY WIND SPEED FORECASTING

OSAMAH BASHEER SHUKUR

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

SEPTEMBER 2015
To my dearest father and mother,
   And
My beloved wife and dear children.
I would like to express my sincere appreciation to my thesis supervisor, Professor Dr. Muhammad Hisyam Lee, for encouragement, guidance, critics and friendship. He supported and directed me throughout my study in UTM and my thesis writing. Without his continued support and interest, this thesis would not have been the same as presented here.
ABSTRACT

The wind speed forecasting is important to observe the wind behaviour in the future and control the harms caused by high or slow speeds. Daily wind speed is more consistent and reliable than other time scales by providing vast monitoring and effective planning. Although a linear autoregressive integrated moving average (ARIMA) model has been used for wind speed forecasting in many recent studies, but the model is unable to identify the nonlinear pattern of wind speed data. ARIMA modelling process causes the stochastic uncertainty as a second reason of inaccurate forecasting results. Wind speed data collection process faces several problems such as the failure of data observing devices or other casual problems that lead losing parts of data. Therefore, wind speed data naturally contains missing values. In this study, an ARIMA-artificial neural network (ANN) and ARIMA-Kalman filter (KF) methods are proposed to improve wind speed forecasting by handling the nonlinearity and the uncertainty respectively. A new hybrid KF-ANN method based on the ARIMA model improves the accuracy of wind speed forecasting by rectifying both nonlinearity and uncertainty jointly. These proposed methods are compared with others such as AR-ANN, AR-KF, and Zhang’s method. AR-ANN method is also used to impute the missing values. It is capable to overcome the missing values problem in wind speed data with nonlinear characteristic. It is compared with linear, nearest neighbour, and state space methods. Two different daily wind speed data from Iraq and Malaysia have been used as case studies. The forecasting results of the ARIMA-ANN, ARIMA-KF and the new hybrid KF-ANN methods have shown in better forecasting than other compared methods, while AR-KF and AR-ANN methods provided acceptable forecasts compared to ARIMA model. The ARIMA-ANN and the new hybrid KF-ANN methods outperformed all other methods. The comparison of missing values imputation methods has shown that AR-ANN outperformed the others. In conclusion, the ARIMA-ANN and the new hybrid KF-ANN can be used to forecast wind speed data with nonlinearity and uncertainty characteristics more accurately. The imputation method AR-ANN can be used to impute the missing values accurately in wind speed data with nonlinear characteristic.
ABSTRAK

TABLE OF CONTENTS

CHAPTER TITLE PAGE

DECLARATION ii
DEDICATION iii
ACKNOWLEDGEMENT iv
ABSTRACT v
ABSTRAK vi
TABLE OF CONTENTS vii
LIST OF TABLES x
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xvi
LIST OF APPENDICES xviii

1 INTRODUCTION 1
  1.1 Background of the Study 1
  1.2 Problem Statement 6
  1.3 Research Question 7
  1.4 Research Objectives 8
  1.5 Significance of the Research 8
  1.6 Scope of the Study 10

2 LITERATURE REVIEW 12
  2.1 Introduction 12
  2.2 Wind Speed Forecasting Methods 14
      2.2.1 ARIMA 14
2.2.2 Artificial neural network 15
2.2.3 Kalman Filter 16
2.2.4 Hybrid Methods 18
2.3 Missing Values Imputation 21

3 METHODOLOGY 23
3.1 Introduction 23
3.2 Time Series Analysis and Forecasting 24
3.2.1 Box-Jenkins Approach 24
3.2.1.1 Model Identification 25
3.2.1.2 Model Estimation 39
3.2.1.3 Diagnostic Checking 40
3.2.1.4 Forecasting 41
3.3 ANN Approach 42
3.3.1 Hybrid ARIMA-ANN Model 47
3.3.2 Hybrid AR-ANN Model 50
3.3.3 Zhang’s Model 51
3.4 KF Approach 53
3.4.1 Hybrid ARIMA-KF Model 55
3.4.2 Hybrid AR-KF Model 59
3.5 Hybrid KF-ANN Model 61
3.6 Forecasting Error Measurements 63
3.7 Hybrid AR-ANN Imputation Method 65
3.8 The Classical Imputation Methods 67

4 RESULTS AND DISCUSSION 69
4.1 Introduction 69
4.2 Iraqi Wind Speed Forecasting 69
4.2.1 ARIMA Model 70
4.2.2 Hybrid AR-ANN Forecasting Method 75
4.2.3 Zhang’s Forecasting Method 77
4.2.4 Hybrid ARIMA-ANN Forecasting Method 78
4.2.5 Hybrid AR-KF Forecasting Method 83
4.2.6 Hybrid ARIMA-KF Forecasting Method 85
4.2.7 Hybrid KF-ANN Forecasting Method 87

4.3 Malaysian Wind Speed Forecasting 92
4.3.1 ARIMA Model 93
4.3.2 Hybrid AR-ANN Forecasting Method 99
4.3.3 Zhang’s Forecasting Method 101
4.3.4 Hybrid ARIMA-ANN Forecasting Method 102
4.3.5 Hybrid AR-KF Forecasting Method 106
4.3.6 Hybrid ARIMA-KF Forecasting Method 108
4.3.7 Hybrid KF-ANN Forecasting Method 109

4.4 Imputing the Missing Values in Iraqi Wind Speed Data 114
4.5 Imputing the Missing Values in Malaysian Wind Speed Data 120
4.6 Summary 125

5 SUMMARY AND CONCLUSIONS 127
5.1 Wind Speed Forecasting Methods 127
5.2 Missing Values Imputation Methods 128
5.3 Recommended Future Studies 129

REFERENCES 130
Appendices A – F 141-175
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>ACF and PACF patterns according to time series models</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>The parameter estimators and the testing values of Iraqi ARIMA model</td>
<td>73</td>
</tr>
<tr>
<td>4.2</td>
<td>MAPE of Iraqi ARIMA forecasts</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>MAPE of Iraqi AR-ANN forecasts</td>
<td>77</td>
</tr>
<tr>
<td>4.4</td>
<td>MAPE of Iraqi Zhang’s forecasts</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>MAPE of Iraqi ARIMA-ANN forecasts</td>
<td>80</td>
</tr>
<tr>
<td>4.6</td>
<td>MAPE of Iraqi AR-KF forecasts</td>
<td>85</td>
</tr>
<tr>
<td>4.7</td>
<td>MAPE of Iraqi ARIMA-KF forecasts</td>
<td>87</td>
</tr>
<tr>
<td>4.8</td>
<td>MAPE of Iraqi KF-ANN forecasts</td>
<td>89</td>
</tr>
<tr>
<td>4.9</td>
<td>The parameter estimates and the testing values of ARIMA</td>
<td>97</td>
</tr>
<tr>
<td>4.10</td>
<td>MAPE of Malaysian ARIMA forecasts</td>
<td>98</td>
</tr>
<tr>
<td>4.11</td>
<td>MAPE of Malaysian AR-ANN forecasts</td>
<td>100</td>
</tr>
<tr>
<td>4.12</td>
<td>MAPE of Malaysian Zhang’s forecasts</td>
<td>101</td>
</tr>
<tr>
<td>4.13</td>
<td>MAPE of Malaysian ARIMA-ANN forecasts</td>
<td>103</td>
</tr>
</tbody>
</table>
4.14 MAPE of Malaysian AR-KF forecasts 107
4.15 MAPE of ARIMA-KF forecasts 109
4.16 MAPE of Malaysian KF-ANN forecasts 111
4.17 Missing datasets of Iraqi wind speed data 114
4.18 RMSE of the estimated series after missing values imputation using the classical methods for all Iraqi missing datasets 115
4.19 RMSE of the estimated series after missing values imputation using hybrid AR-ANN for all Iraqi missing datasets 118
4.20 Missing datasets of Malaysian wind speed data 120
4.21 RMSE of the estimated series after missing values imputation using the classical methods for all Malaysian missing datasets 121
4.22 RMSE of the estimated series after missing values imputation using hybrid AR-ANN for all Malaysian missing datasets 123
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Transfer function types of ANN</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>The general structural of ANN</td>
<td>46</td>
</tr>
<tr>
<td>4.1</td>
<td>Time series plot of Iraqi wind speed</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>ACF and PACF plots of the original wind speed data for Iraq a) ACF, b) PACF</td>
<td>71</td>
</tr>
<tr>
<td>4.3</td>
<td>ACF and PACF plots of the first successive differenced for Iraq a) ACF, b) PACF</td>
<td>72</td>
</tr>
<tr>
<td>4.4</td>
<td>AIC of the AR for Iraqi datasets</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Hybrid ARIMA-ANN training forecasts and the original plots for Iraq</td>
<td>81</td>
</tr>
<tr>
<td>4.6</td>
<td>Hybrid ARIMA-ANN testing forecasts and the original plots for Iraq</td>
<td>81</td>
</tr>
<tr>
<td>4.7</td>
<td>The improvement percentage of training Iraqi forecast after using ARIMA-ANN compared to other conventional methods</td>
<td>82</td>
</tr>
<tr>
<td>4.8</td>
<td>The improvement percentage of testing Iraqi forecast after using ARIMA-ANN compared to other conventional methods</td>
<td>82</td>
</tr>
</tbody>
</table>
methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9</td>
<td>Hybrid KF-ANN training forecasts and the original plots for Iraq</td>
</tr>
<tr>
<td>4.10</td>
<td>Hybrid KF-ANN testing forecasts and the original plots for Iraq</td>
</tr>
<tr>
<td>4.11</td>
<td>The improvement percentage of training Iraqi forecast after using KF-ANN compared to other forecasting methods</td>
</tr>
<tr>
<td>4.12</td>
<td>The improvement percentage of testing Iraqi forecast after using KF-ANN compared to other forecasting methods</td>
</tr>
<tr>
<td>4.13</td>
<td>Time series plot of Malaysian wind speed</td>
</tr>
<tr>
<td>4.14</td>
<td>ACF and PACF plots of the original wind speed data for Malaysia a) ACF, b) PACF</td>
</tr>
<tr>
<td>4.15</td>
<td>ACF and PACF plots of the first successive differenced wind speed data for Malaysia a) ACF, b) PACF</td>
</tr>
<tr>
<td>4.16</td>
<td>ACF and PACF plots of the first successive differenced and the first seasonal differenced wind speed data for Malaysia a) ACF, b) PACF</td>
</tr>
<tr>
<td>4.17</td>
<td>AIC of the AR for Malaysian datasets</td>
</tr>
<tr>
<td>4.18</td>
<td>Hybrid ARIMA-ANN training forecasts and the original plots for Malaysia</td>
</tr>
<tr>
<td>4.19</td>
<td>Hybrid ARIMA-ANN testing forecasts and the original plots for Malaysia</td>
</tr>
<tr>
<td>4.20</td>
<td>The improvement percentage of training Malaysian forecast after using ARIMA-ANN compared to other conventional</td>
</tr>
</tbody>
</table>
methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.21</td>
<td>The improvement percentage of testing Malaysian forecast after using ARIMA-ANN compared to other conventional methods</td>
</tr>
<tr>
<td>4.22</td>
<td>Hybrid KF-ANN training forecasts and the original plots for Malaysia</td>
</tr>
<tr>
<td>4.23</td>
<td>Hybrid KF-ANN testing forecasts and the original plots for Malaysia</td>
</tr>
<tr>
<td>4.24</td>
<td>The improvement percentage of training Malaysian forecast after using KF-ANN compared to other forecasting methods</td>
</tr>
<tr>
<td>4.25</td>
<td>The improvement percentage of testing Malaysian forecast after using KF-ANN compared to other forecasting methods</td>
</tr>
<tr>
<td>4.26</td>
<td>AIC plot for the five datasets of Iraq</td>
</tr>
<tr>
<td>4.27</td>
<td>The improvement percentage of all imputed Iraqi missing datasets using AR-ANN compared to other classical methods</td>
</tr>
<tr>
<td>4.28</td>
<td>AIC plot for the five datasets of Malaysia</td>
</tr>
<tr>
<td>4.29</td>
<td>The improvement percentage of all imputed Malaysian missing datasets using AR-ANN compared to other classical methods</td>
</tr>
<tr>
<td>A.1</td>
<td>Data manager dialog box</td>
</tr>
<tr>
<td>A.2</td>
<td>Import dialog box</td>
</tr>
<tr>
<td>A.3</td>
<td>Neural Network creation</td>
</tr>
<tr>
<td></td>
<td>Description</td>
</tr>
<tr>
<td>---</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>A.4</td>
<td>Training process dialog box</td>
</tr>
<tr>
<td>A.5</td>
<td>Testing process dialog box</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>Autoregressive Integrated Moving Average</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>MA</td>
<td>Moving Average</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>KF</td>
<td>Kalman Filter</td>
</tr>
<tr>
<td>ACF</td>
<td>Autocorrelation Function</td>
</tr>
<tr>
<td>PACF</td>
<td>Partial Autocorrelation Function</td>
</tr>
<tr>
<td>FFBP</td>
<td>Feed-Forward Back Propagation</td>
</tr>
<tr>
<td>LM</td>
<td>Levenberg-Marquardt</td>
</tr>
<tr>
<td>BR</td>
<td>Bayesian Regularization</td>
</tr>
<tr>
<td>SE</td>
<td>State Equation</td>
</tr>
<tr>
<td>OE</td>
<td>Observation Equation</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MMSE</td>
<td>Minimum Mean Square Error</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage Error</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>NWP</td>
<td>Numerical Weather Prediction</td>
</tr>
<tr>
<td>ERNN</td>
<td>Elman's Recurrent Neural Networks</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>EKF</td>
<td>Extended Kalman Filter</td>
</tr>
<tr>
<td>EMD</td>
<td>Empirical Mode of Decomposition</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Name</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>UKF</td>
<td>Unscented Kalman Filter</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-Organizing Map</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The outlines of using ANN toolbox in MATLAB for forecasting</td>
<td>141</td>
</tr>
<tr>
<td>B</td>
<td>Coding for MATLAB M file programming to obtain the forecasting using Kalman filter</td>
<td>147</td>
</tr>
<tr>
<td>C</td>
<td>Coding for MATLAB M file programming to impute the missing values using Linear, nearest neighbour, and state space methods</td>
<td>150</td>
</tr>
<tr>
<td>D</td>
<td>Published and accepted papers</td>
<td>154</td>
</tr>
<tr>
<td>F</td>
<td>Data of Malaysian wind speed</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>Data of Iraqi wind speed</td>
<td>155</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Frequent, extreme wind speeds and the nonlinear nature of wind speed data makes forecasting a complex process. The accuracy of wind speed forecasting is important to secure, control, optimize, and improve renewable wind power generation. The chaotic fluctuation in the pattern of wind speed data is often the reason of the nonlinearity.

Some authors have proposed using autoregressive integrated moving average (ARIMA) model, a classical statistical approach, to forecast wind speeds (Benth and Benth, 2010; Shi et al., 2011). AR, ARIMA, or seasonal ARIMA models have been used for comparing with other methods by (Cadenas and Rivera, 2007, 2010; Chen and Yu, 2014; Guo et al., 2012; Liu et al., 2012b; Tatinati and Veluvolu, 2013). Finding the appropriate wind speed ARIMA model can be accomplished by following the approach proposed by Box-Jenkins. Zhu and Genton (2012) reviewed statistical short-term wind speed forecasting models, including autoregressive models and traditional time series approaches, used in wind power developments to determine which model provided the most accurate forecasts. Although an ARIMA
The nonlinear pattern of wind speed data may be one reason for the forecasting inaccuracy of ARIMA model, which is a linear time series model (Cadenas and Rivera, 2007, 2010). An ANN can be used to handle the nonlinear nature of wind speed data. ANN was proposed to improve the forecasting accuracy of nonlinear time series data by (Assareh et al., 2012; Bilgili and Sahin, 2013; Peng et al., 2013; Pourmousavi-Kani and Ardehali, 2011).

A hybrid ARIMA-ANN model is proposed in this study in order to accommodate all the linear and nonlinear components of wind speed data and combine them in one distinct approach. The proposed hybrid model depends on all ARIMA inputs and their intersections which are those on the right side of ARIMA equation to determine the inputs structure of the ANN and provides the most accurate forecasts.

Several recent studies have proposed different hybrid approaches that combine ARIMA and ANN. Zhang’s hybrid model that combines both ARIMA and ANN models was proposed by (Zhang, 2003) and used by Aladag et al. (2009); and Cadenas and Rivera (2010). Zhang’s hybrid model combined linear and nonlinear components of ARIMA and improved just the residual of ARIMA using ANN.

An ANN can be constructed based on the autoregressive (AR) in order to simulate the ANN structure, this approach can be called hybrid AR-ANN model (Liu et al., 2012b) or only ANN (Khashei and Bijari, 2010, 2011; Zhang, 2003). It was used and compared with other approaches by Guo et al. (2012); Li and Shi (2010); and Liu et al. (2012b). Khashei and Bijari (2010) proposed a hybrid ANN using ARIMA model called an artificial neural network \((p,d,q)\) model.
The inaccurate forecasting of ARIMA model is a problem that reflects the stochastic uncertainty of modelling process as another reason of inaccurate wind speed forecasting results. The Kalman Filter (KF) model can be used for meteorological purposes, such as wind speed forecasting (Cassola and Burlando, 2012; Galanis et al., 2006; Louka et al., 2008). To obtain the best initial parameters for the KF, an ARIMA model is used to create the structure of the KF model that is regarded as the best model for handling the stochastic uncertainty and improve wind speed forecasting. An ARIMA model is used with the KF model to construct the structure of the state equation. This model can be called a hybrid ARIMA-KF. Determining KF state equation structures and ANN inputs structure has been done based on AR, ARIMA, or other time series models by Cadenas and Rivera (2007, 2010); Chen and Yu (2014); Guo et al. (2012); Liu et al. (2012b); Malmberg et al. (2005); Tatinati and Veluvolu (2013); and Zhu and Genton (2012).

In this study, a new hybrid KF-ANN model is proposed based on an ARIMA model to further improve the forecasting accuracy of wind speed. ANN and KF are useful for handling nonlinearity and stochastic uncertainty problems associated with wind speed data. Therefore, ANN and KF improve the accuracy of wind speed forecasting. Many recent studies combined the KF model to handle stochastic uncertainty, with another converged approach, such as support vector machines which handle the nonlinearity of wind speed (Chen and Yu, 2014; Tatinati and Veluvolu, 2013). In the proposed KF-ANN approach, first the KF state (system) and observation (measurement) equations are created based on an ARIMA model. In a second step, the inputs variables of the ANN approaches are generated from the new state series that is the output of the state equation, while the target is the original wind speed series. As a result, the output of the ANN represents the final fitted or forecasting series.

The hybrid ARIMA-ANN and hybrid KF-ANN models resulted in better wind speed forecasting accuracy than their components, while the KF model and ANN separately provided acceptable forecasts compared to ARIMA model that
provided ineffectual forecasts. The hybrid ARIMA-ANN model outperformed all other studied methods.

The collection process of wind speed data as one of meteorological time series data faces several tactical problems such as thunderstorms, failure of data observing devices, or other unforeseen errors that lead to increased complexity in data analysis. A sequential dataset is required for performing analysis and modelling processes. Therefore, the missing values in wind speed data should be filled and imputed. Missing values imputation can be accomplished using simple methods such as linear, nearest neighbour or others. Although complex methods require additional expertise and specialization, they often outperform the simple methods. The classical methods such as linear, nearest neighbour, and state space may not provide accurate imputations when the wind speed data contains nonlinearity.

In most meteorological time series datasets, nonlinearity is a problem that may hamper time series analysis using linear methods as mentioned previously. In particular, wind speed data suffer from nonlinearity in addition to the missing values. In recent studies, ANN was introduced to impute missing values and to handle the nonlinearity of meteorological time series datasets in general and wind speed dataset in particular (Coulibaly and Evora, 2007; He et al., 2013; Junninen et al., 2004; Kim and Pachepsky, 2010; Yozgatligil et al., 2013). Using ANN to impute missing values was not limited to meteorological time series data as Kornelsen and Coulibaly (2012) introduced ANN as the most effective method for missing values infilling in soil moisture as hydrometeorological time series dataset.

In this study, a hybrid AR-ANN method is proposed to reform the problems of this study by imputing missing values and jointly handling the nonlinearity problem. Feed-forward back propagation is used as a neural network algorithm. AR model is used only for determining the structure of the input layer for the ANN. AR order can be determined by observing the significant lags in partial autocorrelation function (PACF). Listwise deletion is also used as the simplest method before AR
modelling to handle missing value problems in wind speed time series datasets. AR-ANN method has been used in many recent papers for handling the nonlinearity in full dataset of wind speed without any missing value as mentioned previously.

The stationarity conditions for the wind speed data were omitted in this stage, and the parameters, signs, and residual series were also omitted from the terms of AR equation, because the AR model was used only for determining the structure of the input layer for the ANN (Khashei and Bijari, 2010; Liu et al., 2012b).

In many studies, AR model was also used purely as missing values imputation method (Aloh, 2009; Choong et al., 2009; Honaker and King, 2010). Applying the listwise deletion was suggested to produce consistent and unbiased attributes of parameters especially for performing data analysis and modelling using model or software that requires a sequential dataset (Cheema, 2014; Honaker and King, 2010).

Many other methods of missing values imputation are compared with AR-ANN method proposed in this study. Linear method and nearest neighbour method are presented as more simple methods for imputing missing values. ANN method and state space method are presented as complex methods that need more expertise and scientific specialization. A linear method is summarized by connecting two data points with a linear equation line. It was used for comparing with more complex methods in many recent papers such as in (Junninen et al., 2004; Kornelsen and Coulibaly, 2012; Norazian et al., 2008). A nearest neighbour method is the simplest imputation method of missing values that can be summarized by replacing the missing values by the nearest neighbour data point. A nearest neighbour method was used as a simple method for comparing with other proposed methods such as in (Junninen et al., 2004; Liew et al., 2011; Siripitayananon et al., 2003; Waljee et al., 2013). The state space model has been proposed and used as an imputation method of missing values in (Sarkka et al., 2004; Tsay, 2005). Root mean square error (RMSE) measurement is computed for the error of missing values imputation for all
imputation methods and all datasets as a statistical criterion to evaluate the adequacy and accuracy of these methods. AR-ANN method has been compared with linear, nearest neighbour, and state space methods. The proposed ANN method outperformed other imputation methods. The results have shown that ANN outperformed the other imputation methods.

In conclusion, the wind speed data with nonlinearity and uncertainty characteristics can be forecasted more accurately using the hybrid models KF-ANN and ARIMA-ANN. The missing values in wind speed data with nonlinear characteristic can be imputed more accurately using a hybrid AR-ANN method.

1.2 Problem Statement

Although an ARIMA model is the preferable statistical model for forecasting, it can lead to inaccurate results for wind speed forecasting. The nonlinear pattern of wind speed data may be one reason for the forecasting inaccuracy of ARIMA model, which is a linear time series model (Cadenas and Rivera, 2007, 2010). It is important to propose an appropriate method to accommodate all the linear and nonlinear components of wind speed data and combine them in one distinct approach in order to handle the nonlinearity.

The inaccurate forecasting of ARIMA model is a problem that reflects the stochastic uncertainty of modelling process as another reason of inaccurate wind speed forecasting results. Proposing a method that handles the stochastic uncertainty problem is important to obtain accurate forecasting. Handling nonlinearity and stochastic uncertainty problems jointly associated with wind speed data using a suitable method is requested for more forecasting accuracy.
The collection process of wind speed data as one of meteorological time series data faces several tactical problems such as thunderstorms, failure of data observing devices, or other unforeseen errors that lead to increased complexity in data analysis. A sequential dataset is required for performing analysis and modelling processes. Therefore, the missing values in wind speed data should be filled and imputed. Missing values imputation can be accomplished using the classical methods such as linear, nearest neighbour, and others. The classical methods may not provide accurate imputations when the wind speed data contains nonlinearity.

1.3 Research Question

(a) What are the methods proposed in this study to improve the forecasting accuracy?

(b) What are the most appropriate forecasting methods among the classical and the hybrid methods for Iraqi and Malaysian wind speed data?

(c) What are the methods used to impute the missing values in Iraqi and Malaysian wind speed data?

(d) What is the most appropriate method for missing values imputation among the classical and the proposed methods?
1.4 Research Objectives

The following objectives are recognized to achieve the aims of the research:

(a) To develop ARIMA-ANN hybrid model to handle the nonlinearity in wind speeds and to develop ARIMA-KF hybrid model to handle the stochastic uncertainty.

(b) To develop a new KF-ANN hybrid model to handle nonlinearity and stochastic uncertainty problems jointly those associated with wind speed data.

(c) To evaluate the forecasting performance of classical and proposed methods.

(d) To develop hybrid AR-ANN model to impute the missing values in wind speed data.

(e) To evaluate the performance of classical and proposed methods for imputing the missing values in wind speed datasets.

1.5 Significance of the Research

Some authors have proposed using ARIMA model, a classical statistical approach, to forecast wind speeds. It can lead to inaccurate results for wind speed forecasting. The nonlinear pattern of wind speed data may be one reason for the
forecasting inaccuracy of ARIMA model, which is a linear time series model (Cadenas and Rivera, 2007, 2010). An ANN can be used to handle the nonlinear nature of wind speed data. A hybrid ARIMA-ANN model is proposed in this study in order to accommodate all the linear and nonlinear components of wind speed data and combine them in one distinct approach for more handling of the nonlinearity. The hybrid ARIMA-ANN model depends on all ARIMA inputs and their intersections which are those on the right side of ARIMA equation to determine the inputs structure of the ANN and provides the most accurate forecasts. Several recent studies have proposed different hybrid approaches that also combine ARIMA and ANN. Zhang’s hybrid model that combines both ARIMA and ANN was proposed by (Zhang, 2003) and used by Aladag et al. (2009); and Cadenas and Rivera (2010). Zhang’s hybrid model only improved the residuals of ARIMA by using ANN. Constructing an ANN based on AR to simulate the ANN structure was used and compared with other approaches by Guo et al. (2012); Li and Shi (2010); and Liu et al. (2012b). The advantages of the new hybrid ARIMA-ANN model were hybridizing the linear ARIMA model and a nonlinear ANN and combining their components in one distinct approach to improve wind speed forecasting accuracy.

The inaccurate forecasting of ARIMA model reflects the stochastic uncertainty of modelling process as another reason of inaccurate wind speed forecasting. An ARIMA model is used with the KF model to construct the structure of the state-space equation. In recent studies, researchers proposed using hybrid AR-KF model instead of hybrid ARIMA-KF model to maintain simplicity when the parameters of moving average and integration parts become zero (Chen and Yu, 2014; Liu et al., 2012b; Tatinati and Veluvolu, 2013). In this study, the KF model is initialized based on ARIMA(p,d,q)(P,D,Q)s to obtain a hybrid ARIMA-KF model. ARIMA-KF model combines all the components of both ARIMA and KF in distinct approach for more forecasting accuracy than AR-KF model.

In this study, a new hybrid KF-ANN model is proposed based on an ARIMA model as a unique method for handling the nonlinearity and uncertainty jointly to further improve the forecasting accuracy of wind speed. The wind speed data with
nonlinearity and uncertainty characteristics can be forecasted more accurately using the hybrid models KF-ANN and ARIMA-ANN.

A sequential dataset is required for performing analysis and modelling processes. Missing values imputation can be accomplished using classical methods such as linear, nearest neighbour or others. The classical methods may not provide accurate imputations when the wind speed data contains nonlinearity. In this study, a hybrid AR-ANN model was proposed to reform the missing values problem by imputing missing values and jointly handling the nonlinearity problem. This method can be called hybrid AR-ANN model (Liu et al., 2012b) or ANN (Khashei and Bijari, 2010, 2011; Zhang, 2003). The missing values in wind speed data with nonlinear characteristic can be imputed more accurately using AR-ANN model.

1.6 Scope of the Study

In this study, daily wind speed data from two meteorological stations were collected. The first dataset was collected from the Mosul dam meteorological station in Mosul, Iraq ¹. It covered four hydrological years (1 October 2000 – 30 September 2004) which was used for training. Another four months of hydrological data (1 October 2004 – 31 January 2005) was reserved for testing. The other dataset was collected from the Muar meteorological station in Johor, Malaysia ². It covered four hydrological years (1 October 2006 – 30 September 2010) which was used for training. An additional three months of hydrological data (1 October 2010 – 31 December 2010) was used for testing. These samples of data are applied to test the proposed methods. Appendix E and Appendix F include full Iraqi and Malaysian wind speed datasets respectively.

¹ Dam and Water Resources Research Center, University of Mosul, Mosul, Iraq.
² Departments of Environment, Ministry of Natural Resources and environment, Malaysia.
Missing values datasets have been generated by distributing the missing values into the training periods of wind speed datasets. The missing values have been distributed into three different parts 1, 3, 6 randomly. Three different proportions 10%, 20%, 30% have been considered percentages of missing values. Five datasets is the total number of missing datasets for each of Iraq and Malaysia. First three datasets include 10% of missing values that were distributed into one, three, and six equal parts respectively. 10% requires advanced methods for handling as mentioned in (Acuña and Rodriguez, 2004). Last two datasets were distributed into six equal parts that includes 20%, and 30% of missing values respectively. Acuña and Rodriguez (2004) mentioned that any percentage of missing values more than 15% may impact the scientific interpretation.
REFERENCES


