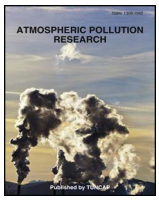
Atmospheric Pollution Research 9 (2018) 358–368

HOSTED BY

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr



Application of computational intelligence techniques to forecast daily PM_{10} exceedances in Brunei Darussalam

Sam-Quarcoo Dotse[a,*], Mohammad Iskandar Petra[a], Lalit Dagar[b], Liyanage C. De Silva[a]

[a] Department of Systems Engineering, Faculty of Integrated Technologies, Universiti Brunei Darussalam, Jalan Tungku Link, BE1410, Brunei Darussalam
[b] Environmental Studies, Faculty of Arts and Social Sciences, Universiti Brunei Darussalam, Jalan Tungku Link, BE1410, Brunei Darussalam

ARTICLE INFO

Keywords:
PM_{10}
Artificial neural networks
Genetic algorithm
Random forests
Variable selection
Brunei Darussalam

ABSTRACT

Particulate matter (PM_{10}) is the pollutant causing exceedances of ambient air quality thresholds, and the key indicator of air quality index in Brunei Darussalam for haze related episodes caused by the recurrent biomass fires in Southeast Asia. The present study aims at providing suitable forecasts for PM_{10} exceedances to aid in health advisory during haze episodes at the four administrative districts of the country. A framework based on random forests (RFs), genetic algorithm (GA) and back propagation neural networks (BPNN) computational intelligence techniques has been proposed in which the final prediction is made by the BPNN model. A hybrid combination of GA and RFs is initially applied to determine optimal set of inputs from the initial data sets of largely available meteorological, persistency of high pollution levels, short and long term variations of emissions rates parameters. The inputs selection procedure does not depend on the back propagation training algorithm. The numerical results presented in this paper show that the proposed model not only produced satisfactory forecasts but also consistently performed better via several statistical performance indicators when compared with the standard BPNN and GA optimisation based on back propagation training algorithm. The model also showed satisfactory threshold exceedances forecasts achieving for instance best true predicted rate of 0.800, false positive rate of 0.014, false alarm rate of 0.333 and success index of 0.786 at Brunei-Muara district monitoring station. Overall, the current study has profound implications on future studies to develop a real-time air quality forecasting system to support haze management.

1. Introduction

Brunei Darussalam has in recent years experienced haze conditions which range from slight transient haze episodes to severe haze episodes largely due to the long-range transport of pollutants from biomass fires in Southeast Asia (SEA) during the regular dry seasons. Particulate matter (PM_{10}) emitted as a result of the biomass fires is the pollutant causing exceedances of ambient air quality thresholds and the key indicator of air quality index in the country (Dotse et al., 2016a). Several scientific studies have linked many adverse effects of both short-term and long-term exposures to ambient particles on human health (WHO, 2013). The 1997-98 SEA haze episodes considered to be the worst air pollution incidents on record in the country has for example been linked to incidence of respiratory diseases (Anaman and Ibrahim, 2003; Yadav et al., 2003). Volatile organic compounds (VOCs) and heavy metals, some of which are known or suspected carcinogens, mutagens, and teratogens, which have the potential to cause serious long-term effects were characterised in the 1998 haze episode (Muraleedharan et al., 2000). The recurrent haze episodes have become one of the top environmental concerns in the country due to the potential effect on human health and the environment. The Government established the National Haze Action Plan to safeguard the health and safety of the public through the prevention and mitigation of land and forest fires, and control emissions, including the prohibition of open burning during the regular dry period (http://www.env.gov.bn/). There is also an updated national emission inventory of greenhouse gases and criteria pollutants based on government statistics and other sources to help to assess air quality management programs (Dotse et al., 2016b). A recent study on the temporal and spatial distribution of PM_{10} based on a long-term monitoring data and the factors that influence high particulate matter events have been conducted (Dotse et al., 2016a). There are other earlier studies into the sources and characteristics limited to 1997-98 haze episodes in the country (Muraleedharan et al., 2000; Radojevic and Hassan, 1999; Radojevic, 2003). Nevertheless, air pollutants forecasting is an important component of any air quality control and management system such as the National Haze Action Plan. A real-time air quality forecasting system is crucial to obtain advance knowledge on whether the

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.
* Corresponding author. Tel.: +673 8998952.
E-mail addresses: 14h0302@ubd.edu.bn, samdotse@yahoo.com (S.-Q. Dotse).

http://dx.doi.org/10.1016/j.apr.2017.11.004
Received 1 August 2017; Received in revised form 1 November 2017; Accepted 5 November 2017
Available online 08 November 2017

pollutant concentrations would exceed the given guidelines or limit values in the country provided by Ministry of Health and the department of Environment, Parks and Recreations for health advisory during haze episodes (MOH, 2013).

Significant studies have in recent years been devoted to improve statistical models for air quality forecasting as they rely mainly on routinely available historical data, and are therefore considered affordable and easy to implement (eg. Díaz-Robles et al., 2008; Ul-Saufie et al., 2013). They are generally more suitable for the description of complex site-specific relations between concentrations of air pollutants and potential predictors, and often have a higher accuracy, as compared to deterministic models (Zhang et al., 2012). Non-linear statistical models which stemmed from different machine learning algorithms for regression tasks in the field of Computational Intelligence (CI) have been successfully applied to air quality forecasting problems. Artificial Neural Networks (or simply neural networks or NN) CI models have attracted a large amount of attention among the statistical approaches due to several advantages. NN model can model highly non-linear functions and can be trained to accurately generalize when presented with new, unseen data; also, unlike other statistical techniques it makes no prior assumptions concerning the data distribution (Gardner and Dorling, 1998). Despite the attractiveness of neural networks models, the design of the best networks architecture and the choice of optimal input variables still remain a major challenge to its predictive performance. Experimental results showed that hybrid models can effectively improve NN forecasting accuracy obtained by either of the models used separately (Díaz-Robles et al., 2008). In most cases, the other models or methods combined with neural networks are usually used to determine the optimal inputs parameters and consequently the best networks architecture which significantly enhance the forecast accuracy. The selection of input variables is therefore a key issue, since irrelevant or noisy variables may have negative effects on the training process, resulting in an unnecessarily complex model structure and poor generalization power (Voukantsis et al., 2011). Fewer input variables reduce the complexity of the model (Sousa et al., 2007). Whereas there are various types of variable selection methods, many of the automated model selection methods, such as backward or forward stepwise regression, are classical solutions to this problem, but are generally based on strong assumptions about the functional form of the model or the distribution of residuals (Sandri and Zuccolotto, 2006). It is on this basis that a framework based on artificial neural networks, genetic algorithm (GA) and random forests (RFs) CI techniques is proposed to forecast PM_{10} exceedances in this study. The input variables selection is done using a hybrid model that combines GA and RFs learning algorithms as a single algorithm in which GA controls the selection process. This 'wrapper' variable selection method utilises RFs learning algorithm as a black box to score subsets of variables from the initial data set according to their predictive power. The main advantage of selecting relevant variables through an algorithmic modeling technique is the independence from any assumptions on the relationships among variables and on the distribution of errors (Sandri and Zuccolotto, 2006). The various input variable selection methods with their advantages and disadvantages are discussed in Kohavi and John (1997) and Guyon and Elisseeff (2003).

Genetic Algorithms are stochastic search strategies developed as the inspiration of biologic evolution and have been successfully used to solve different optimisation problems in wide range of application areas (Scrucca, 2013). As an evolutionary algorithms guided by several parameters, the fitness function is an important parameter that controls the selection and survival of each individual at each generation. Previous application of GAs to select input variables selection in forecasting particulate matter concentrations using neural networks mostly utilised the networks learning algorithms as fitness function in the optimisation process (Niska et al., 2004; Grivas and Chaloulakou, 2006; Antanasijević et al., 2013). This work considers random forests fitness function. The RF algorithm is based on an aggregation of many binary decision trees obtained using the classification and regression trees (CART) method (Breiman, 2001; Breiman et al., 1984), and makes use of bagging (bootstrap aggregation) to combines multiple random predictors in order to aggregate predictions (Brence and Brown, 2006) allowing for high complexity without over-generalising and over-fitting to the training data (Ho, 1995). In view of this, a number of RFs could be drawn from a larger RF forming an initial population of individuals; genetic algorithms could be an ideal optimisation solution to build a more accurate ensemble (Bader-El-Den and Gaber, 2012). RF is primarily for prediction but capable of ranking the input variables in terms of their importance to the model. It is increasingly being used for input variable selection due to the many advantages over other learning algorithms, and has been effectively used to select inputs in air quality forecasting problems (Jollois et al., 2009; Poggi and Portier, 2011). Though its application has seen growing popularity in many disciplines, very limited literature exists in the field of air quality modeling to the best of our knowledge.

Artificial neural networks have previously been applied to forecast particulate matter in some major cities (eg Díaz-Robles et al., 2008; Voukantsis et al., 2011; Ul-Saufie et al., 2013). The main aim of this study is to apply multi-layer back propagation neural networks to obtain suitable forecasts of daily peaks of PM_{10} concentrations at four air quality monitoring stations across Brunei Darussalam, in support of the National Haze Action Plan. The inputs for the prediction is taken from meteorological, persistency of high pollution levels, short and long term variations of emissions rates parameters. Local meteorology plays an important role in the day-to-day variations of PM_{10} concentrations and its seasonality across Brunei Darussalam (Dotse et al., 2016a). Airborne particles exhibits diurnal variation, typically rising through the night to very high levels in the early morning and thereafter decreases due largely to meteorological factors (Radojevic and Hassan, 1999). Meteorological variables are important inputs to in developing any prediction model for the country. There are however largely available meteorological variables at different averaging times and the complex interactions between them, and therefore the need for an effective procedure to select the most significant variables. A hybrid model that combines GA and RFs is therefore applied to select optimal set of inputs from the initial data sets before the final neural network prediction model. The numerical results of the proposed framework are compared with genetic algorithm input variables optimisation based on back propagation training algorithm, and the standard back propagation neural networks models.

2. Materials and methods

2.1. Study location and data

Brunei Darussalam (Latitude 4.8903°N, Longitude 114.9422°E) with an area of 5765 sq. km and a population of 393,372 in 2011 is made up of four districts: Brunei-Muara, Tutong, Belait and Temburong. The capital is Bandar Seri Begawan (BSB), located in Brunei-Muara District, which is the smallest and the most densely populated district. The districts of Brunei-Muara, Tutong and Belait, which form the larger western portion, are dominated by hilly lowlands, swampy plains and alluvial valleys. Mountainous terrain abounds in the eastern district of Temburong. The climate in the country is generally hot and wet throughout the year. The main sources of particulate matter pollution are the transboundary haze episodes in Southeast Asia and occasionally localised fires in Brunei and in neighbouring Malaysian states of Sabah and Sarawak Dotse et al., 2016a; Radojevic and Hassan, 1999). The low wind system of the country coupled with the hilly lowlands, swampy plains and alluvial valleys topographic features of some parts do not favor the dispersion of air pollutants but instead bring in pollutants into the country. Transboundary pollution from industrial centers, forest fires, and volcanic eruptions in other countries in the region can also have significant effects on particulate matter in Brunei. Emissions from

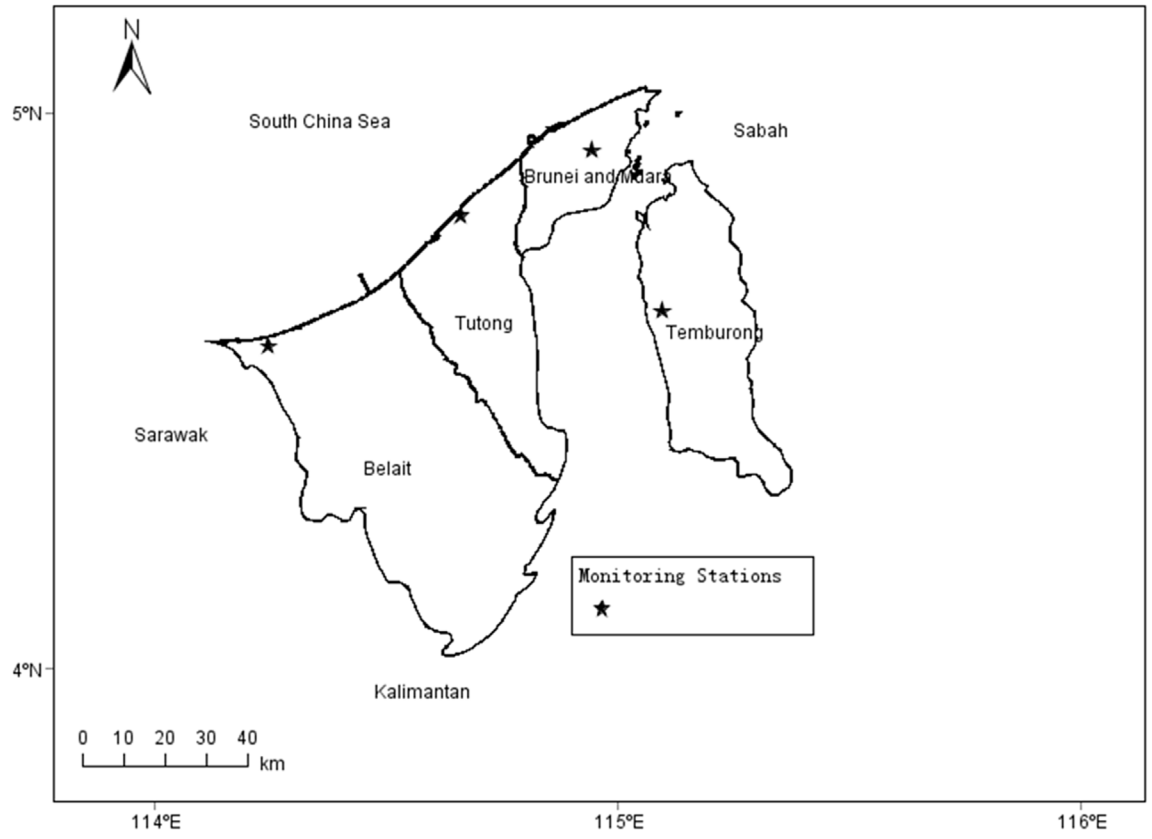


Fig. 1. Map of Brunei Darussalam showing the four administrative districts and locations of the air quality monitoring stations.

Table 1
The descriptive statistics of PM_{10} concentrations at the four monitoring stations (2009–2013).

District	Brunei-Muara					Temburong				
	2009	2010	2011	2012	2013	2009	2010	2011	2012	2013
Data availability (days)	265	352	365	366	365	272	355	365	366	365
Mean	15.51	9.98	10.92	11.01	11.32	24.13	15.16	15.75	17.85	16.01
Median	12.00	9.00	8.00	8.70	9.70	19.00	14.00	15.00	15.80	14.20
Minimum	2.00	2.00	1.00	2.60	3.10	4.00	4.00	5.00	7.00	8.10
Maximum	77.00	38.00	67.00	36.00	80.90	127.00	37.00	52.00	49.40	98.30
Standard deviation	11.89	4.83	8.15	6.71	7.54	19.31	5.89	7.15	7.88	9.33
Variance	141.39	23.36	66.43	45.03	56.82	372.92	34.70	51.08	62.06	87.05
Skewness	2.65	1.73	3.26	1.63	4.80	3.03	1.21	2.28	1.51	4.96
Kurtosis	11.66	7.59	17.54	5.12	35.23	14.06	4.40	10.82	5.38	36.29
Number of days exceeded 50 μgm^{-3}	8	–	4	–	4	12	–	4	–	5
Annual average	15	10	10	11	11	19	15	16	18	16

District	Tutong					Belait				
	2009	2010	2011	2012	2013	2009	2010	2011	2012	2013
Data availability (days)	332	352	365	366	365	272	362	365	366	365
Mean	21.22	15.77	19.20	20.12	19.55	24.13	17.42	22.66	20.79	25.23
Median	17.00	15.00	17.00	17.60	16.90	19.00	16.00	21.00	18.70	21.40
Minimum	5.00	3.00	5.00	4.10	5.00	4.00	3.00	6.00	1.50	3.40
Maximum	194.00	39.00	82.00	55.20	123.00	127.00	51.00	93.00	62.60	101.40
Standard deviation	17.35	5.42	9.14	9.66	11.66	19.31	6.41	9.48	9.09	14.30
Variance	300.93	29.40	83.51	93.23	135.92	372.92	41.03	89.80	82.68	204.54
Skewness	6.65	1.00	2.56	1.61	4.74	3.03	1.45	2.51	1.77	2.68
Kurtosis	61.40	4.40	13.99	5.43	34.89	14.06	7.15	14.62	7.26	1230
Number of days exceeded 50 μgm^{-3}	8	–	5	7	6	21	2	8	6	22
Annual average	22	16	19	20	20	21	17	23	21	25

motor vehicles, and industrial processes and solvent use, have been identified as potential sources of particulate matter in the country (Dotse et al., 2016b). Pollution Control Division of the Department of Environment, Parks and Recreation currently maintains and operates networks of air quality monitoring stations located throughout the four administrative districts. A five year daily mean PM_{10} concentration data

(2009–2013) from four monitoring stations each located in the four districts is used in the study. Fig. 1 is the map of the country showing the four administrative districts and locations of the air quality monitoring stations. Table 1 also gives the descriptive statistics of PM_{10} at the four locations.

The meteorological data used in the study is provided by the Brunei Darussalam Meteorological Department, under the Ministry of Communications. It was not possible to obtain up to date meteorological data at the air quality monitoring sites for the study period. Therefore, the meteorological record at Brunei International Airport (BIA) is used in the analysis. BIA is located in Brunei-Muara districts and the data from station is considered to be representative of the atmospheric conditions, and also taking into account the country size and the locations of the air quality monitoring stations. Several meteorological variables were analysed in order to determine those that influence daily PM_{10} concentration across the country through correlation analysis. A Spearman's rank correlation coefficient was used as the PM_{10} values were not distributed normally. The initial set of meteorological model inputs were selected based on those that were found to correlate significantly with daily PM_{10} concentration and they include: daily rainfall (Rain), temperature difference (T_{diff}) minimum, maximum and mean values of temperature (T_{min}, T_{max} and T_{av}) and relative humidity (RH_{min}, RH_{max} and RH_{av}), highest and mean wind speed (WS_{max} and WS_{av}), and wind direction (WD). Sine and cosine transformations were employed for the Wind Direction in order to replace its cyclic nature with a linear one as in equation (1) and similar to Karatzas and Kaltsatos, 2007 and Voukantsis et al., 2011.

$$\mathrm{Sin(WD)} = \frac{\mathrm{Sin}(2\pi(\mathrm{x}-\mathrm{min(x)}))}{\mathrm{max(x)}-\mathrm{min(x)}},\quad \mathrm{Cos(WD)} = \frac{\mathrm{Cos}(2\pi(\mathrm{x}-\mathrm{min(x)}))}{\mathrm{max(x)}-\mathrm{min(x)}},\quad \mathrm{x}\in[0^0, 360^0] \tag{1}$$

where sin(WD) and cos(WD) are the results of the linearised wind direction (WD), calculated as a function of the sine and the cosine of the difference of the WD from the minimum value monitored, divided by the difference between the maximum and minimum WD values that have been monitored. The descriptive statistics meteorological input variables and their correlation with PM_{10} concentration at the four locations is presented in Table 2. Biomass burning activities linked to Southeast Asian regional haze episodes have been identified as the main source of high PM_{10} concentrations in the country. Notwithstanding, a recent emission inventory of greenhouse gases and criteria pollutants based on government statistics and other sources revealed contributions from industrial processes and solvent, road transport and power plants to particulate matter emissions (Dotse et al., 2016b). Emissions from motor vehicles could be a major future source of particulate matter due to growing vehicle fleet. It is therefore important to account for the difference in PM_{10} emissions between weekdays and weekends in our prediction models. The daily PM_{10} concentrations also showed clear patterns of seasonal variations across the country with the highest concentrations recorded during the southwest monsoon months from June to September. Therefore, month of the year (MOY) input parameter is considered to accounts for the seasonality in PM_{10} in the model. The effect of day of the week (DOW) and MOY parameters were considered using a suitable arithmetic index, in order to account for the short-term variability in the intensity of emission sources (see Ziomas et al., 1995). Numbers 1 through 7 are assigned to Sunday through Saturday in the day-of-week attribute, and 1 through 12 are assigned to January through December for month of the year attribute. As in the case of wind direction, sine and cosine variables were generated for DOW and MOY parameters (Voukantsis et al., 2011). The possibility of occurrence of pollution episodes is increased if the previous day's pollution levels were higher than normal (Ziomas et al., 1995). Previous day daily PM_{10} concentration ($LagPM_{10}$) is therefore used as input to account for the persistency of high pollution levels in the atmosphere.

The PM_{10} and meteorological data sets were preprocessed. The missing values in the datasets were replaced by multiple imputation approach using Expectation Maximization Based (EMB) algorithm. All missing values imputations were done on R platform using Amelia II package (Honaker et al., 2011).

An appropriate data formulation is required in order to establish the domain knowledge for effectively training of the intelligent system to forecast PM_{10} exceedances. As explained in section 3.1, only data matrices for March and June–September during the five year period were therefore used in the simulation. The available data is then divided into two subsets. Four year data (2009–2012) was used for training the NN models, in which a portion (20%) was used for cross validation during the training process. The second subset, data for the year 2013 was used as testing set to evaluate the trained models. This is important to avoid overfitting. The training set is used to estimate the model parameters, the validation set to choose among a set of different already trained alternative models, and the testing set to run the chosen approximating function on previously unseen data, in order to get an objective measure of its generalisation performances (Corani, 2005). It is important to note that the extreme values (outliers) detected in preprocessing process were also not removed but used to create objective training data sets that would enable the final models to generalise well with extreme PM_{10} values (exceedances). Finally, the data sets were normalised to a similar magnitude in the range of $[-1, 1]$ for the neural networks implementation.

2.2. Methods

The computational analysis was carried out in the R environment (www.r-project.org). There are two main procedures involved in the methodology: 1) Genetic algorithm optimisation scheme to select input parameters, and 2) the training of neural networks to obtain the final model. Genetic algorithm simulates the evolution of living organisms, where the fittest individuals dominate over the weaker ones, by mimicking the biological mechanisms of evolution, such as selection, crossover and mutation (Scrucca, 2013). The standard GA algorithm which consists of population, selection, crossover and mutation has been adopted in this study. Briefly: the GA modeling process begins with a randomly generated population of individuals (chromosomes), which are the possible solutions to the problem with each one of these individuals having a chance of being selected to generate the next offspring. The algorithm then evaluates the fitness of each individual and only the fittest individuals reproduce, passing their genetic information to their offspring. The process is iterated through a sequence of successive generations by implementing genetic search operators (crossover and mutations) until an optimal solution is obtained according to the given stopping criterions from the fitness function. Random forests (RF) is the fitness function in this study. It is a hybrid scheme that combines GA and random forests as a single algorithm in which GA controls the variable selection process by optimising RF tuning parameters. Random forests is a very efficient algorithm based on an aggregation of many binary decision trees obtained using the classification and regression trees (CART) method (Breiman et al., 1984). In terms of variable selection, the initial input variables are ranked in terms of their importance to the prediction model. The number of input variables randomly chosen at each split, mtry and the number of trees in the forest, ntree are the two main tuning parameters to be optimise in the genetic algorithm procedure. To evaluate the quality of the fitted model, the error is estimated through the Out-Of-Bag (OOB) error, calculated according to the iterations of the algorithm. The OOB error corresponds to the prediction error for the data not belonging to the bootstrap sample used to build the tree, which explains its name. Detailed theoretical development can be found in Breiman (2001). Random forests implementation in R is based on randomForest package (Liaw and Wiener, 2002). The genetic algorithm optimisation procedure was carried out using the caret package in R (Kuhn, 2008) which has a mechanism to check overfitting using internal and external performance estimates of the fitted random forests model depending on

Table 2
(a) The descriptive statistics of meteorological input variables, and (b) their correlation with PM10 concentrations at the four locations.

(a)

Meteorological variables[a]	Minimum	Maximum	1st Quartile	3rd Quartile	Mean	Median	Standard deviation
T_{av} (°C)	23.40	30.40	27.00	28.40	27.67	27.70	1.00
T_{min} (°C)	20.60	26.80	23.50	24.60	24.06	24.00	0.81
T_{max} (°C)	26.00	37.60	31.40	33.10	32.25	32.30	1.35
T_{diff} (°C)	2.50	14.20	7.30	9.10	8.19	8.20	1.34
RH_{av} (%)	63.00	97.00	80.00	86.00	82.86	83.00	4.62
RH_{min} (%)	20.00	91.00	57.00	68.00	62.33	63.00	7.84
RH_{max} (%)	79.00	100.00	95.00	98.00	96.41	97.00	2.53
WS_{av} (m/s)	0.51	7.20	2.06	2.73	2.46	2.37	0.59
WS_{max} (m/s)	3.60	19.03	6.17	8.75	7.72	7.20	2.01
Rain (mm)	0.00	195.10	0.00	9.80	9.43	0.80	19.18

(b)

Meteorological variables	PM_{10} concentration (μgm^{-3})			
	Brunei-Muara	Belait	Tutong	Temburong
T_{av} (°C)	0.4146	0.3321	0.3144	0.3861
T_{min} (°C)	0.1634	0.1644	0.1423	0.1599
T_{max} (°C)	0.3979	0.2720	0.2774	0.3902
T_{diff} (°C)	0.2872	0.1696	0.1872	0.2874
RH_{av} (%)	−0.4904	−0.4011	−0.3945	−0.4574
RH_{min} (%)	−0.3676	−0.3128	−0.2921	−0.3857
RH_{max} (%)	−0.5126	−0.3954	−0.4403	−0.4498
WS_{av} (m/s)	−0.1591	−0.0313	−0.0985	−0.1250
WS_{max} (m/s)	−0.0854	0.0028	−0.0154	−0.0338
Rain (mm)	0.4249	−0.4059	−0.3861	−0.3813

[a] See section 2.1 for variable descriptions.

the resampling strategy. The GA implementation in caret uses the underlying code from the GA package (Scrucca, 2013).

The final part of the methodology involves training neural networks using the optimal inputs determined in GA application. Multilayer perceptron (MLP) with learning based on error back propagation is most successful and widely used architecture air quality forecasting due to its accuracy and reliability. MLP network architecture consists of a system of interconnected information processing units called neurons or nodes which are arranged in layers, namely input, hidden, and the output layers (Rumelhart et al., 1986; Mishra and Goyal, 2015). The nodes are connected by weights and output signals which are a function of the sum of the inputs to the node modified by a simple nonlinear transfer, or activation function (Gardner and Dorling, 1998). The final model used in this study is a three-layer feedforward back propagation network type. The number of neurons in hidden layer has a strong influence on the output because too few neurons will contribute to underfitting, while too many neurons lead to over-fitting (Ul-Saufie et al., 2013). However, there are no reliable guidelines to determine the number of neurons in the hidden layer as the appropriate number depends on many factors, including number of input and output neurons, the amount of training data, the amount of noisy data, and the complexity of the learning task. The number of the hidden layer nodes were tested from 1 to 30 and the optimal network configurations obtained after repeated computations. Different numbers of hidden neurons were obtained for the best performance models as different numbers of inputs variables selected for at the four stations. The selected best activation function for both input and hidden layer was hyperbolic tangent sigmoid function which ranges from −1 to 1. The data sets were therefore normalised to a similar magnitude in the range of [−1, 1]. The training of NN models was done using neuralnet package in R (Günther and Fritsch, 2010). In accordance with the two main procedures and the flow chart in Fig. 2, the following steps were followed to obtain final model to forecast of daily PM_{10} exceedances at four monitoring stations across Brunei Darussalam: (i) Divide the data into a training and test datasets, (ii) run GA – RF optimisation procedure on the training data sets to select input variables (GA_{RF}), (iii) train the neural networks (BPNN) using the inputs selected on the training data, and (iv) evaluate the optimal trained models performances using the test data set to obtain the final GA_{RF}-BPNN model. The procedure (i) – (iv) is repeated by using back propagation training algorithm in step (ii) to genetically optimised back propagation neural networks (GA_{BP}-BPNN). Regarding the GA parameters, the initial population was 50, and the initial weights and thresholds were normalised in the range of [-1, 1] for the BP algorithm. The crossover probability was 0.8 and mutation probability was 0.1. A trained BPNN based on the initial inputs is also produced. The proposed framework of genetic algorithm, random forests, and neural networks is now compared with GA_{BP}-BPNN and BPNN.

2.3. Evaluation of the models

The model performance during training and validation processes is assessed with several statistical performance measures (equations (2)–(6)) that are frequently used in the field of air quality forecasting (Voukantsis et al., 2011; Antanasijević et al., 2013; Ul-Saufie et al., 2013). Let O_i represent observed and P_i the predicted values and their respective mean values as $\overline{O}$ and $\overline{P}$. σ represents the standard deviation of the sample data set.

- The correlation coefficient (r)

$$r = \frac{\overline{(O_i - \overline{O}_i)\cdot(P_i - \overline{P}_i)}}{\sigma_{P_i}\cdot\sigma_{P_o}} \quad (2)$$

r is a dimensionless indicator ranging from −1 to 1 that reflects the extent of a linear relationship between the observed and the predicted values.

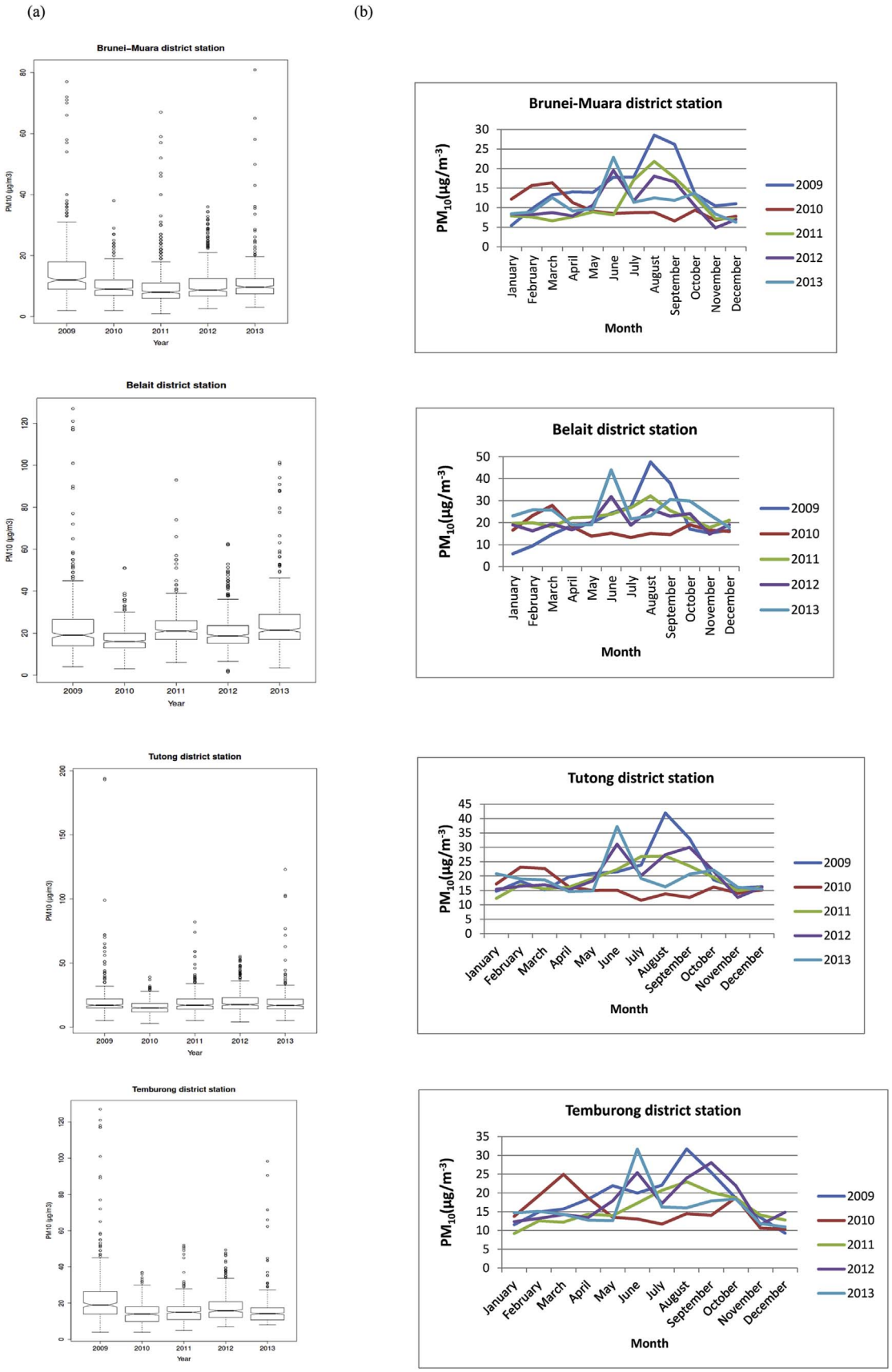


Fig. 2. (a). Boxplots of daily PM_{10} concentrations at the four monitoring stations (2009–2013), **(b)**. Monthly variations of daily PM_{10} concentrations at the four monitoring stations.

- The index of agreement (IA)

$$IA = 1 - \frac{\sum_i^N (O_i - P_i)^2}{\sum_i^N (|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \quad (3)$$

IA is limited to the range 0–1, with values closer to 1 indicating good agreement between observed and predicted values.

- The mean absolute error (MAE)

$$MAE = \frac{\sum_i |O_i - P_i|}{N} \quad (4)$$

MAE value closer to zero indicates good agreement between observed and predicted values.

- The root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum_i^N (O_i - P_i)^2}{N}} \quad (5)$$

RMSE is a measure of the total deviation of predicted values from observed values.

- The mean bias error (MBE)

$$MBE = \overline{P_i} - \overline{O_i} \quad (6)$$

Mean bias error (MBE) defines whether a model over- (positive value) or under- (negative value) predicts the observations.

In addition to these statistical measures, the true predicted rate (TPR), the false positive rate (FPR), the false alarm rate (FAR) and the success index (SI) are used to investigate the model ability to forecast daily PM_{10} concentrations threshold exceedances (Corani, 2005). Given that C represents observed and correctly predicted exceedances, O for all observed exceedances, P for all predicted exceedances and N for the total observations, TPR = C/O, FPR = (P – C)/(N-O), FAR = (P – A)/P and SI = TPR – FPR.

3. Results and discussions

3.1. Overview of the PM_{10} concentrations data

A brief overview of the relevant pollution characteristics of the five year PM_{10} data used in this study which is taken from monitoring stations each located in the four administrative districts of Brunei Darussalam. The decision to focus on data for March, and June–September in the modeling and simulation is based on the temporal and spatial distributions of PM_{10} across the country. Fig. 2 (a) gives the graphical depiction of the daily PM_{10} concentrations for each year by boxplots. Each of the notched boxplots depicts the median (middle line of the box, notches represent the upper and lower 95 percent confidence interval), upper (25%) and lower (75%) quartile (top and bottom lines of the box, respectively), minimum and maximum values (upper and lower end of the whisker lines). Outliers and extreme PM_{10} values are shown by points outside the whisker line which are recorded at all stations and for every year. The mean, median, skewness and kurtosis values are presented in Table 1. For each year the mean were higher than the median values indicating that the data were skewed to the right and the occurrence of high or extreme PM_{10} values. This is also indicated by the high positive skewness and kurtosis values. The kurtosis for each year at all stations show that the data is not normally distributed. It is clear from the descriptive statistics of the PM_{10} values measured at the four locations for the period under consideration as presented in Table 1 and illustrated in Fig. 2 (a) that they were occurrences of high episodes. The air quality index is based on the principle of the Pollution Standard Index (PSI) used by the US Environmental Protection Agency (USEPA). As started earlier in the introduction PM_{10} is the pollutant causing exceedances of ambient air quality thresholds and the key indicator of air quality index in the country. During haze episodes in Brunei, PSI is invariably based on PM_{10}, as the concentrations greatly exceed those of other criteria pollutants, SO_2, CO, NO_2 and O_3 (Radojevic and Hassan, 1999). The daily exceedances of concentration greater than 50 μgm^{-3} guideline limit established by the Brunei Darussalam Ministry of Health (MOH, 2013) and the department of Environment, Parks and Recreations for health advisory during haze episodes is included in Table 1. Air quality is considered good, and outdoor activities are allowed for all age groups as air pollution poses little or no risk when the PSI reading is below 50 μgm^{-3}. PM_{10} concentration is usually below the USEPA, European Union (EU) and World Health Organisation standards for most part of the year with the lowest concentrations occurring at monitoring sites located in Brunei-Muara and Temburong districts. The analysis revealed the daily exceedances almost occurred within the southwest monsoon months of June to September linked to SEA haze episodes. The monthly variations of daily PM_{10} concentrations in Fig. 2 (b) showed high peaks for in March, and June–September. The high peaks are due to long-range transport of smoke particles from the agricultural biomass burning and forest fires in northern SEA countries during the northeast monsoon. The wet northeast monsoon season in southern SEA region is characterised by a dry season in these parts of the region. Also, the occasional localised fires in Brunei, and in the nearby border regions of Sarawak and Sabah are also linked to high values in March. Further discussion on the temporal and spatial distributions PM_{10} and the influence of SEA episodes is reported in Dotse et al. (2016a). The inter-annual variations can be observed from the boxplots and the annual averages are also provided in Table 1.

3.2. The genetically optimised random forests – back propagation neural networks

The numerical results of the genetically optimised random forests – back propagation neural networks (GA_{RF}-BPNN) applied to forecast daily PM_{10} exceedances at the four monitoring stations are presented in this section. As stated in the methodology, in order to train BPNN to obtain the final prediction model, genetic algorithm based on random forests fitting function applied to the initial data sets in order to determine optimal set of model inputs. The selected inputs variables for each station by GA optimisation procedure are given in Table 3. The most common and relevant variables selected at the four locations are previous day PM_{10} ($LagPM_{10}$), rainfall, wind speed and the month of year. Temperature and relative humidity were also selected though not common for all four stations. It is worth noting that these variables are sufficient in explaining the underlying mechanism behind the transport

Table 3
The selected inputs variables for each station for the genetically optimised random forests.

Input variables	Brunei Muara	Belait	Tutong	Temburong
Previous day PM_{10}	✓	✓	✓	✓
Rainfall	✓	✓	✓	✓
Mean temperature		✓		
Minimum temperature				
Maximum temperature				
Temperature difference				
Minimum relative humidity			✓	
Maximum relative humidity				
Mean relative humidity				
Maximum wind speed	✓			✓
Mean wind speed	✓		✓	
Sine of the wind direction				
Cosine of the wind direction				
Sine of the day of week				✓
Cosine of the day of week		✓		✓
Sine of the month of year	✓	✓		
Cosine of the month of year			✓	

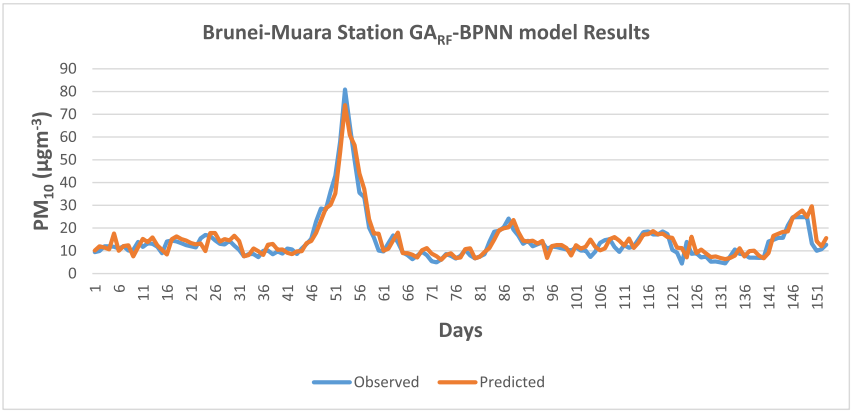
and distributions of PM_{10} concentrations across the country and the occurrence of high episodes. The day-to-day variation in daily PM_{10} across the country is determined by temperature, relative humidity and rainfall. High episodes are usually associated with high temperatures and low amounts of rainfall and relative humidity. In addition, wind speed and direction also play an important role in the occurrences of high episodes and are mainly responsible for its seasonality. Therefore according to the selected inputs variables, the month of year and wind speed variables could account for the seasonal variation, whilst rainfall, temperature and relative humidity inputs account for the for daily variation in PM_{10} concentrations in the predictive models. Similarly, $LagPM_{10}$ and wind speed variables could account for high peaks. The selected inputs are now used to train the neural networks to obtain the final model to forecast daily mean PM_{10} concentrations for the next day at each location. The optimum trained GA_{RF}-BPNN architecture is selected based on four statistical performance indicators including, linear correlation coefficient (r), index of agreement (IA), mean absolute error (MAE) and root mean square error (RMSE) (see section 2.3). The r and IA are used to check the accuracy of the model result; values closer to 1 indicate higher accuracy. Whereas, RMSE and MAE are used to quantify the error in model; a value closer to 0 indicates good performance. As indicated in section 2.2., four year data (2009–2012) which is 80% was used for building the NN models (training and validation) and one year (2013) which is 20% for testing in order to avoid overfitting. The validation set is about 20% of the training sets. The model has performed well during the training and validation processes. The numerical results of performance indicators presented in the study are based on a comparison of the best model results for the test data sets with actual observations. The values of r, IA, RMSE and MAE are 0.9502, 0.9727, 3.2942 μgm^{-3} and 2.4032 μgm^{-3} respectively for Brunei-Muara district station, 0.9397, 0.9677, 4.5346 $\mu gm^{-3}$and 3.1072 μgm^{-3} for Temburong, 0.9112, 0.9079, 10.3299 μgm^{-3} and 7.5557 μgm^{-3} for Belait, and 0.8725, 0.8451, 11.0044 μgm^{-3} and 8.2211 μgm^{-3} for Tutong district station. These results obtained for the proposed hybrid model at the four stations were generally satisfactory. PM_{10} time series plots of the observed and predicted values (March, June–September 2013) at the four selected locations are also presented in Fig. 3.

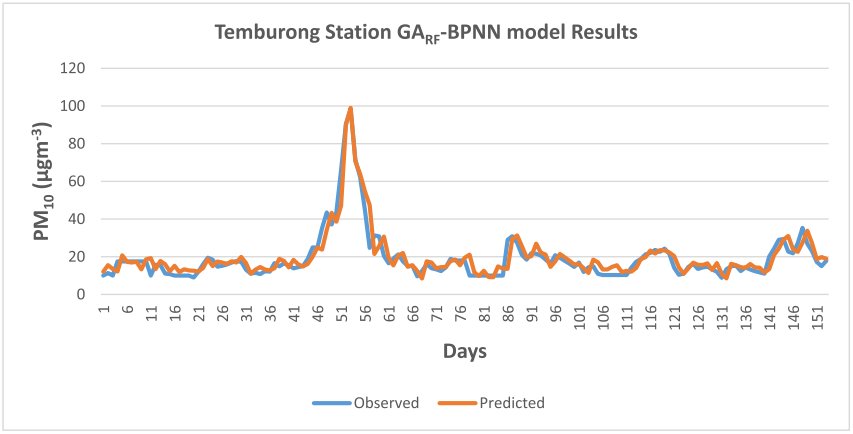
The main aim of adopting the random forests, genetic algorithm and neural networks framework in this study is to achieve more accurate forecasts of daily PM_{10} exceedances to aid in health advisory during haze episodes in the four districts of Brunei Darussalam. The standard back propagation neural networks (BPNN) trained with all seventeen input variables and a genetic algorithm optimisation of input variables based on back propagation training algorithm (GA_{BP}-BPNN) models were constructed to investigate whether there is any significant improvement in the approach in GA_{RF}-BPNN model. Similar steps were followed in training and evaluation of the BPNN and GA_{BP}-BPNN models using the same training, validation and testing data sets for each monitoring station. Table 4 gives the statistical performance indicators of all the three models which are calculated by comparing the model results for the test data sets with actual observations. Though, all the three models performed satisfactory, it is evident from the table that there is significant improvement in the forecasts produced by the proposed GA_{RF}-BPNN model compared with the standard BPNN model. Also, the GA_{BP}-BPNN model performed better that the BPNN model at the four stations. The two hybrid models (GA_{RF}-BPNN and GA_{BP}-BPNN) performed better than BPNN because the GA optimisation procedure ensured that variables that have little or no effect on the predictive performance of the backpropagation neural networks were removed from the model's inputs. This enhanced the hybrid models accuracies by reducing the complexity networks, the running time and the uncertainties associated with generalisation. The GA_{RF}-BPNN proposed model performed slightly better than GA_{BP}-BPNN model. However, training time of GA_{RF}-BPNN model significantly reduced and the errors generally lower than GA_{BP}-BPNN model.

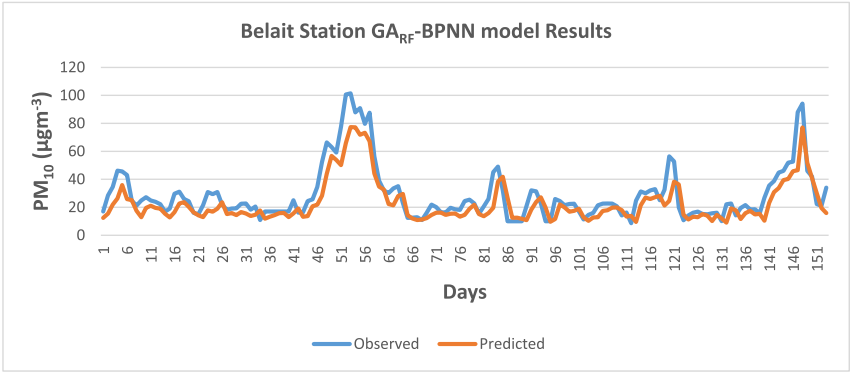
The ultimate goal is to have a model able to accurately forecast daily peaks at which a decision can be made to issue an alarm on exceedances. Therefore the ability of the models to accurately forecast days where PM_{10} concentrations exceed a given threshold value at the stations is investigated using the true predicted rate (TPR), the false positive rate (FPR), the false alarm rate (FAR) and the success index (SI). TPR determines the fraction of correctly predicted exceedances over total exceedances with values from 0 to 1, FPR is the fraction of false predictions over total non-exceedances with values from 0 to 1, and FAR is the fraction of false predictions over total exceedances with values from 0 to 1. SI determines the fraction of correct predictions over total predictions with values from 0 to 1. The optimum or best model is achieved for TPR and SI close to 1, and FPR and FAR close to zero. As mentioned in section 3.1, the threshold value of PM_{10} concentrations for health advisory during haze episodes is 50 μgm^{-3} but due to the smaller number of exceedances at the stations (see Table 1), a 40 μgm^{-3} threshold has been used in this study in order to effectively evaluate the models based on the threshold indicators. The sensitivity of the detection of the 40 μgm^{-3} exceedances based threshold indicators of the proposed model which performed better than the other two models are included in Table 4. These were calculated by considering a model uncertainty corresponding to the RMSE errors. The TPR and SI values at the Brunei-Muara station are 0.800 and 0.786 respectively, and that of Temburong station are 0.750 and 0.736. The FPR and FAR values for these stations are respectively 0.014 and 0.333 for Brunei-Muara and 0.014 and 0.250 for Temburong. These results show a very good predictive accuracy of PM_{10} exceedances forecasts for the next day at the two locations. The model performed averagely at Belait station with TPR and SI values of 0.500 and 0.460 respectively and FPR and FAR values of 0.04 and 0.263 respectively. However, though the model performance at Tutong station is satisfactory reflected in the earlier statistical indexes (r, IA, RMSE and MAE), and also having FPR value of 0.007 FPR and 0.250 for FAR, the results per TPR and SI values showed that the model performed badly in forecasting exceedances at this location. Also, included in Table 4 for all the models is the mean bias error (MBE) given by the differences in the mean of the observed and predicted PM_{10} concentrations and it defines whether a model over- (positive value) or under- (negative value) predicts the observations. The models only slightly under-predict in some cases. The models only slightly overestimated the daily PM_{10} for the next day at the Brunei-Muara and Temburong stations. However, all three models underestimated the daily values for the next day at Belait and Tutong stations.

In general, the proposed model overall performance in forecasting daily peaks for the next day based on the numerical results presented so far is satisfactory. It has predicted with high accuracy PM_{10} exceedances at Brunei-Muara and Temburong compared with the Belait and Tutong districts stations. The mean reason for the poor performance of the model at Belait and Tutong districts may be probably due to the pollution characteristics of the station locations.

Southeast Asian biomass is the main source of high PM_{10} concentrations in the country which depend largely depends on the prevailing weather conditions, hotspot locations and extent of the fires. The stations located in Belait and Tutong districts have recorded the highest number of exceedances due to their proximity to the regular fire hotspots and topographic features described in section 2.1. Belait district also has peatlands which during dry season is very susceptible to fires thereby leading to PM_{10} exceedances. Another reason may be due to the distance of the meteorological station to the air quality station locations in these areas since some meteorological variables were used as inputs to the predictive models. It was not possible to obtain up to date meteorological variables at the respective stations during the time period which could improve the forecasts. It is important to state that neural networks will fail to extrapolate on data that have not been presented during the training procedure (Gardner and Dorling, 1998). Further investigations are needed in order to increase the forecasting accuracy of the model forecasts at Belait and Tutong district locations.







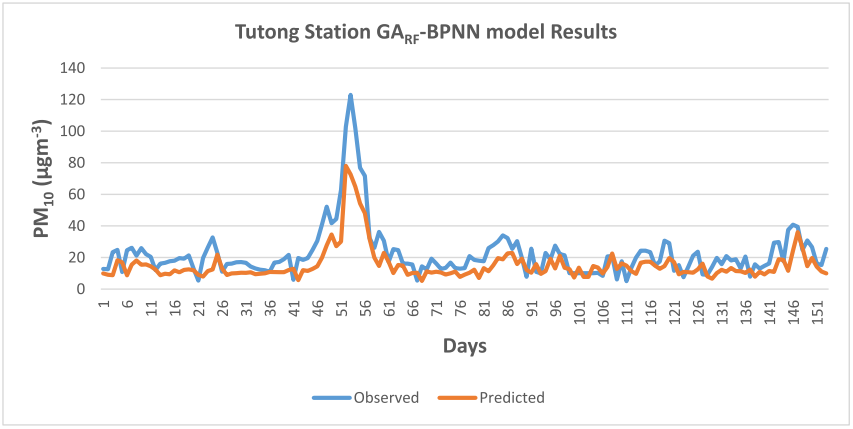


Fig. 3. PM_{10} concentration time series of the observed and predicted values (March, June–September 2013) at the four selected locations based on the proposed model.

This study will be extended in future to include more sources (e.g Sea salts, peatland, oil and gas industry etc) and mechanisms (atmospheric physics and chemistry) related to Haze episodes in the region with the aim of integrating the statistical model with deterministic models which could improve accuracy of predicting PM_{10} concentrations thresholds across the country.

Table 4
Performance statistics for the validation of the developed models at the four locations.

	r	IA	RMSE	MAE	MBE
Brunei-Muara station					
BPNN	0.9223	0.9563	4.0057	2.6297	−0.2330
GA_{BP}-BPNN	0.9266	0.9612	3.9080	2.4246	0.3460
GA_{RF}-BPNN	0.9502	0.9727	3.2942	2.4032	0.7000
Temburong station					
BPNN	0.9072	0.9468	5.5392	3.7586	0.6600
GA_{BP}-BPNN	0.9113	0.9490	5.4059	3.3016	0.3740
GA_{RF}-BPNN	0.9397	0.9677	4.5346	3.1072	0.7800
Belait station					
BPNN	0.8856	0.9026	10.7192	7.8175	−5.6100
GA_{BP}-BPNN	0.9041	0.8934	10.9258	7.9464	−6.3400
GA_{RF}-BPNN	0.9112	0.9079	10.3299	7.5557	−5.9320
Tutong station					
BPNN	0.8302	0.7663	12.8392	9.2527	−7.7940
GA_{BP}-BPNN	0.8589	0.8086	12.0540	8.7824	−7.5800
GA_{RF}-BPNN	0.8726	0.8451	11.0044	8.2211	−7.0500

Threshold Indicators for the proposed model

GA_{RF}-BPNN	TPR	FPR	FAR	SI
Brunei-Muara	0.800	0.014	0.333	0.786
Temburong	0.750	0.014	0.250	0.736
Belait	0.500	0.040	0.263	0.460
Tutong	0.250	0.007	0.250	0.243

4. Conclusions

A framework based on back propagation neural networks (BPNN), genetic algorithm (GA) and random forests (RFs) computational intelligence techniques has been investigated to obtain suitable forecasts for PM_{10} exceedances to aid in health advisory during haze episodes at the four administrative districts of Brunei Darussalam. BPNN formed the final prediction model whereas a hybrid combination of GA and RFs is initially applied to determine optimal set of inputs from the initial data sets of largely available meteorological, persistency of high pollution levels, short and long term variations of emissions rates parameters. Several statistical performance measures frequently used in the field of air quality forecasting were used to assess the model performance during training and validation processes. Also, the ability of the model to accurately forecast days where PM_{10} concentrations exceed a given threshold value at the stations is investigated using the true predicted rate, the false positive rate, the false alarm rate, and the success index threshold indicators. The numerical results presented in this paper show that the proposed genetically optimised random forests – back propagation neural networks prediction model produced satisfactory forecasts daily exceedances for the next day. There was improvement in the forecasts when compared with the numerical results of genetic algorithm optimisation of input variables based on back propagation training algorithm and the standard back propagation neural networks. The model also showed satisfactory threshold exceedances forecasts achieving for instance best true predicted rate of 0.800, the false positive rate of 0.014, the false alarm rate of 0.333 and the success index of 0.786 at Brunei-Muara district monitoring station.

Though satisfactory, the model has predicted with high accuracy PM_{10} exceedances at Brunei-Muara and Temburong compared with the Belait and Tutong districts stations. Overall, the current study has profound implications on future studies to develop a real-time air quality forecasting system to support haze management in Brunei Darussalam and it also highlighted the importance of variable selection in identifying the optimal functional forms of statistical models.

Acknowledgments

The authors gratefully acknowledge the Brunei Darussalam Department of Environments, Parks and Recreation (JASTRE) and the Brunei Darussalam Meteorological Department for respectively providing PM_{10} concentrations and meteorological variable datasets. We would also like to acknowledge the support and Graduate Research Scholarship (GRS) funding provided by the Universiti Brunei Darussalam for this research.

References

Anaman, K.A., Ibrahim, N., 2003. Statistical estimation of dose-response functions of respiratory diseases and societal costs of haze-related air pollution in Brunei Darussalam. Pure Appl. Geophys. 160, 279–293.
Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ., Perić-Grujić, A.A., 2013. PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci. Total Environ. 443, 511–519.
Bader-El-Den, M., Gaber, M., 2012. Garf: towards self-optimised random forests. In: Neural Information Processing. Springer Berlin, Heidelberg, pp. 506–515.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J.H., Ohlsen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Belmont.
Brence, M.J.R., Brown, D.E., 2006. Improving the Robust Random Forest Regression Algorithm. Systems and Information Engineering Technical Papers. Department of Systems and Information Engineering, University of Virginia.
Corani, G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 185 (2), 513–529.
Dotse, S.Q., Dagar, L., Petra, M.I., De Silva, L.C., 2016a. Influence of Southeast Asian haze episodes on high PM 10 concentrations across Brunei Darussalam. Environ. Pollut. 219, 337–352.
Dotse, S.Q., Dagar, L., Petra, M.I., De Silva, L.C., 2016b. Evaluation of national emissions inventories of anthropogenic air pollutants for Brunei Darussalam. Atmos. Environ. 133, 81–92.
Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. 42 (35), 8331–8340.
Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32 (14), 2627–2636.
Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of PM 10 hourly concentrations, in the Greater Area of Athens, Greece. Atmos. Environ. 40 (7), 1216–1229.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Günther, F., Fritsch, S., 2010. Neuralnet: training of neural networks. R. J. 2 (1), 30–38.
Ho, T.K., 1995. Random decision forests. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 1. IEEE, pp. 278–282.
Honaker, J., King, G., Blackwell, M., 2011. Amelia II: a program for missing data. J. Stat. Softw. 45 (7), 1–47.
Jollois, F.X., Poggi, J.M., Portier, B., 2009. Three nonlinear statistical methods to analyze PM_{10} pollution in Rouen area. Case Stud. Bus. Ind. Gov. Stat. 3, 1–17.
Karatzas, K.D., Kaltsatos, S., 2007. Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simul. Model. Pract. Theory 15 (10), 1310–1319.
Kuhn, M., 2008. Caret package. J. Stat. Softw. 28 (5), 1–26.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1–2), 273–324.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. news 2 (3), 18–22.
Mishra, D., Goyal, P., 2015. Development of artificial intelligence based NO_2 forecasting models at Taj Mahal, Agra. Atmos. Pollut. Res. 6 (1), 99–106.
MOH, 2013. Brunei Darussalam Ministry of Health. Health Advisory during Haze Period. www.moh.gov.bn/SiteCollectionDocuments/Haze/healthadvisory- 2013.pdf, Accessed date: 13 August 2016.
Muraleedharan, T.R., Radojevic, M., Waugh, A., Caruana, A., 2000. Chemical characterisation of the haze in Brunei Darussalam during the 1998 episode. Atmos. Environ. 34, 2725–2731.
Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M., 2004. Evolving the neural network model for forecasting air pollution time series. Eng. Appl. Artif. Intell. 17 (2), 159–167.
Poggi, J.M., Portier, B., 2011. PM_{10} forecasting using clusterwise regression. Atmos. Environ. 45 (38), 7005–7014.
Radojevic, M., Hassan, H., 1999. Air quality in Brunei Darussalam during the 1998 haze episode. Atmos. Environ. 33 (22), 3651–3658.
Radojevic, M., 2003. Haze research in Brunei Darussalam during the 1998 episode. Pure Appl. Geophys. 160 (2003), 251–264 0033 – 4553/03/020251 – 14.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representation by back-propagation errors. In: Rumelhart, D.E., McClelland, J.L., The PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, MA.

Sandri, M., Zuccolotto, P., 2006. Variable selection using random forests. In: Data Analysis, Classification and the Forward Search. Springer Berlin, Heidelberg, pp. 263–270.

Scrucca, L., 2013. GA: a package for genetic algorithms in R. J. Stat. Softw. 53 (4), 1–37.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Model. Softw. 22 (1), 97–103.

Ul-Saufie, A.Z., Yahaya, A.S., Ramli, N.A., Rosaida, N., Hamid, H.A., 2013. Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmos. Environ. 77, 621–630.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. 409 (7), 1266–1276.

WHO, 2013. Review of Evidence on Health Aspects of Air Pollution - REVIHAAP Project. World Health Organisation, WHO Regional Office for Europe, Copenhagen.

Yadav, A.K., Kuamr, K., Kasim, M., Singh, M.P., Parida, S.K., Sharan, M., 2003. Visibility and incidence of respiratory diseases during the 1998 haze episode in Brunei Darussalam. Pure Appl. Geophys. 160, 265–277.

Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F., Paliatsos, A.G., 1995. Forecasting peak pollutant levels from meteorological variables. Atmos. Environ. 29 (24), 3703–3711.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60, 632–655.