Engng.J.CMU.(2018) 25 (1), 203-213

วารสารวิศวกรรมศาสตร์
มหาวิทยาลัยเชียงใหม่
Engineering Journal
Chiang Mai University

The Hybrid Neural Networks-ARIMA/X Models and ANFIS Model for PM-10 Forecasting: A Case Study of Chiang Mai, Thailand's High Season

RATI Wongsathan
Department of Electrical and Computer Engineering and
the Research Institute of North-Chiang Mai University
169 Hangdong Chiang Mai 50230, Thailand Tel. +66(81)2893400, Fax. +66(53)819998
E-mail: rati1003@gmail.com, rati@northcm.ac.th

ABSTRACT

It was demonstrated that efficient air quality models are very useful tools in forecasting air pollutants. Consequently, in this paper, the influence of exogenous variable (X) related meteorological parameters and correlated toxic gas together with the significant historical PM-10 values was analyzed to formulate the numerical hybrid PM-10 forecast model during high season (January-April) in Chiang Mai Province, northern Thailand. The hybrid model is divided into two stages, dealing firstly with nonlinear transformation through the multilayer perceptron neural network (MLPNN) and radial basis function neural network (RBFNN), and secondly with statistical estimation of the linearly stationary residuals using an autoregressive integrated moving average (ARIMA) and ARIMAX, and denoted by *h*MLPNN/RBFNN-ARIMA/X model. On the tradeoff between the accuracy using root mean square error (RMSE) and mean absolute error (MAE), and the reliability through Akaike information criterion(AIC) for PM-10 forecasting, the *h*MLPNN-ARIMA model was identified as the optimal model whereas the *h*MLPNN-ARIMAX and the *h*RBFNN-ARIMA model were identified as the sub-optimal model. For further comparison against PM-10 forecast based an adaptive neuro-fuzzy inference system (ANFIS) model, it was indicated that the *h*MLPNN-ARIMA model provided slightly more accurate but clearly more reliable than that of the ANFIS including the others.

Keywords: ARIMA model, ARIMAX model, Multilayer perceptron neural networks, Radial basis function neural network, ANFIS, PM-10.

1. Introduction

Nowadays, the situation of air pollution related to particulate matter less than 10 micrometer in diameter (PM-10) is critical severity for the health of people in upper provinces of northern Thailand. Chiang Mai, the largest city in the north and second-largest city in Thailand, continuously experiences with the pollution related to PM-10. During the high season (around January-April), PM-10 level frequently exceeds the standard level of 120 μg/m^3 specified by Thai government pollution control department (PCD). PM-10's main sources are forest fire and biomass burning. For example only Mae Chaem, one of totally 25 districts of Chiang Mai, produces and burns over 37,000 tons of corncob waste every year. Moreover, Chiang Mai is located in Chiang Mai-Lamphun basin, where smoke from neighboring Myanmar and Lao is prone to settle. The statistical of daily PM-10 behavior at the high season during 2012-2017 monitored by two stations including the provincial hall station (the country side) and the Yuparaj Wittayalai school station (in the city moat area) indicated that the PM-10 level has reached up to 300 μg/m^3 (Fig. 1) whereas in Bangkok, the capital city, the PM-10 level is only 40-50 μg/m^3. In 2014, PM-10 level has exceeded the threshold level and closely reached to the level of 300 μg/m^3 resulting in thousands of people across Chiang Mai area being admitted to hospital for various respiratory illness. In 2015, the PM-10 situation was severely seen by touching of 300 μg/m^3 level but also decreased in 2016 because of the climate variability and off-seasonal rains. Therefore, the existing environment may directly affect to PM-10 level.

Since the PM-l0 data is usually measured and officially announced in the daily morning to warn the people but this information may not be thoroughly accessible. Consequently, people

Received 15 September 2017
Accepted 1 February 2018

cannot prepare to prevent themselves in advance. The temporal PM-10 forecast model should be implemented together with weather forecast in order to minimize health risks to the public. Previously, most of the PM-10 forecast models frequently use the historical PM-10 value to estimate the PM-10 including multivariate linear regression (MLR) [1] as the linear model, and Neural Networks (NN) [2] as the nonlinear model. In general, NNs performance is better than MLR, but MLR is often employed more than NNs due to its simplicity. Besides, other techniques have extensively been applied and made the comparison among them. Some of these techniques are hybrid MLR and ARIMA [2], ARIMA [3]-[4], hybrid ARIMA and NNs [3], support vector machines (SVM) [5], hybrid ARIMA and SVM [6] and etc.

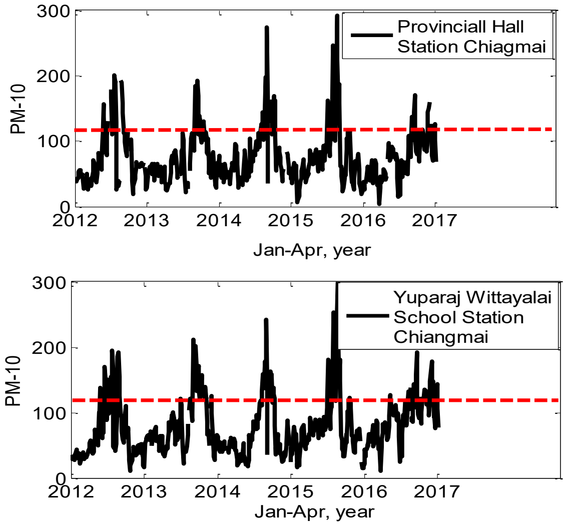

Fig.1 PM-10 time-series of Chiang Mai, January-April 2012-2017.

In the case of PM-10 researches in Chiang Mai, most of them focused on monitoring device implementation which gives more budget support funds. While, existing researches on PM-10 forecasting model still used the conventional regression techniques (e.g., simple linear regression (SLR), MLR [7], grey system model [8]), which were easy to formulate, but the results were unsatisfied. However, the author acknowledges no related research on PM-10 forecasting using nonlinear model or hybrid model except for our previous works [9]-[12] in the series. Early works on PM-10 forecast model using various NNs model such as MLPNN, RBFNN and hybrid of RBFNN and genetic algorithm have already implemented [9]. Overall, they provided accurate forecasting results. Unfortunately, an over-fitting obtained from a learning process is the main disadvantage for NNs as well as the local-trap parameters due to the large structure of the networks, leading to a huge erroneous forecast. To improve the forecast accuracy, *h*ARIMA-NN model was alternately selected and employed to investigate this issue [10]. The forecasting results were demonstrated that it surpassed NNs and ARIMA, respectively. However, the forecast error is considerably high during the high season because of high disturbance and PM-10 variances itself. In this case, the *h*RBFNN-ARIMA model and *h*ARIMA-RBFNN model were proposed to tackle this problem [11]. The *h*RBFNN-ARIMA model was identified the best forecast model in this case. To improve the performance of the forecasting model further, the exogenous variable and the modified hybrid algorithm should be considered. Two hybrid models i.e., *h*ARIMAX-RBFNN model and *h*RBFNN-ARIMAX model were implemented [12]. The forecast results demonstrated that they outperformed the previous models [12] suggesting that the nonlinear model should be firstly captured the non-stationary non-linear component of PM-10 pattern, and the fully linearly stationary residuals can be accurately predicted by the linear model later.

Recently, the hybrid learning of NN and fuzzy inference system or ANFIS was identified as the most accurate models compared to the fuzzy logic model [13]. For PM-10 forecast, it was considered as an efficient tool due to its high accuracy [14]-[15]. However, the most suitable forecast model requires not only for its accuracy but also reliability. The aim of this work is to investigate the accuracy and simultaneous reliability of the hybrid PM-10 forecast models and to compare among them. The different *h*MLPNN/RBFNN-ARIMA/X forecast models were optimized through the experimental design. The ANFIS model was subsequently implemented for PM-10 forecast to compare with them. The forecasting performance were evaluated by tradeoff between accuracy and reliability through the RMSE, MAE and AIC.

2. Methodologyand Method

Data was collected during 2011-2016 for both PM-10 and exogenous variables includes 4 related toxic gas variables (*CO*, O_3, NO_2, and SO_2) and 4 significant meteorological variables (*GW*, *T*, *P*, and *H*) provided by the Thailand PCD. The descriptive statistics results of these variables are presented in Table 1. They are divided into 3 parts i.e., training, validating and testing by using the data from 2011-2014, 2015 and 2016, respectively. Four different PM-10 forecast models including **Model (1)** *h*MLPNN-ARIMA, **Model (2)** *h*RBFNN-ARIMA, **Model (3)** *h*MLPNN-ARIMAX and **Model (4)** *h*RBFN N-ARIMAX were implemented and investigated. Their forecast results were compared to existing forecast models. The model implementations was developed using the MATLAB program and all simulations were run on a 2.27 GHz Intel Pentium Core i5 processor with 6 GB of RAM laptop computer. The description of the proposed forecast models was detailed in the next Sub-Section.

Table 1 The descriptive statistics of 602 testing data of PM-10 and exogenous variables.

Measured Parameter	Symbol	Unit	Range	Min.	Max.	Mean	Variance	Relative Change
PM-10	*PM*	µg/m^3	260.1	29.9	290	42.65	1065	4.09
Carbon monoxide	*CO*	ppm	1.9	0	1.9	0.49	0.08	0.04
Ozone	O_3	ppb	59	5	64	25.68	129.5	2.19
Nitrogen dioxide	NO_2	ppb	33.6	0.2	33.8	9.43	28.18	0.83
Sulfur dioxide	SO_2	ppb	16	0	16	0.73	1.0	0.06
Wind gust	*GW*	km/hr	54.7	0	54.7	21.0	44.8	0.81
Temperature	*T*	Celsius	19.8	19.4	39.2	26.7	8.5	0.43
Pressure	*P*	hPa	18.4	964.4	982.8	973.5	11.8	0.64
Relative Humidity	*H*	-	68	24	92	64.9	136.6	2.0

2.1 The hNN-ARIMA model

The structure of NN model denoted by NN(*P*, *Q*, 1) where *P* is the number of input nodes of the historical PM-10, PM-10$_{t-i}$, i = 1, 2, …, *P*, *Q* is the number of hidden nodes and “1” refers to single output node (Fig. 2).

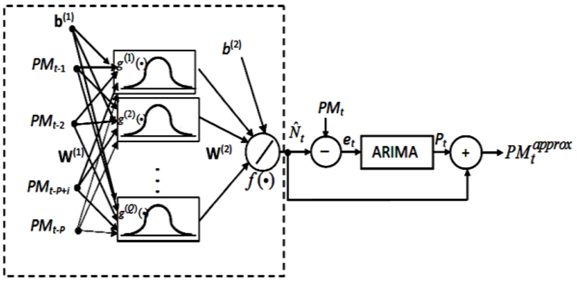

Fig.2 The PM-10 forecast based on *h*NN-ARIMA model.

The nonlinear solution, N_t, is filtered through the NN in the first stage. The error generated at time *t*, e_t, is passed through the linear model in the second stage which is statistically generated through an ARIMA procedure. Once, the correlation of residuals from this model is removed, the *h*NN-ARIMA model is established. Two different types of this model i.e., the *h*MLPNN-ARIMA and the *h*RBFNN-ARIMA model are detailed in the next Sub-Section 2.1.1 and 2.1.2, respectively.

2.1.1 The hMLPNN-ARIMA model

In the first stage, the MLPNN(*P*, *Q*, 1) determines N_t which is expressed in matrix-vector form as,

$$N_t = f\left(\mathbf{W}^{(2)} \times g\left(\mathbf{W}^{(1)} \times \mathbf{PM} + \mathbf{b}^{(1)}\right) + b^{(2)}\right), \qquad (1)$$

where **PM** is the *P*-column input vector of historical PM-10 data, $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ is the *Q*-by-*P* weight matrix and the *Q*-column bias vector between input-hidden layer, respectively, and $\mathbf{W}^{(2)}$ and $b^{(2)}$ is the *Q*-row vector and the bias value between hidden-output layer, respectively. From the test, the sigmoid function yielded the lowest RMSE and was selected as an activation function denoted by *g*, and *f* referred to a linear transfer function. The parameter of MLPNN, including weights and biases, are tuned by the well-known back-propagation algorithm. The results from the optimization showed that the number of input and hidden nodes equals 1. Then, MLPNN (1, 1, 1) is mathematically expressed explicitly as,

$$N_t = \left(\frac{2 \times 0.59}{1+\exp((3.88 \times PM_{t-1} + 1.64))} - 0.59 \right) - 0.41. \quad (2)$$

The error (e) resulted from (2) is passed to ARIMA model in the second stage. A practical approach to construct the ARIMA model includes three iterative steps i.e., identification, parameter estimation and diagnostic checking. For the model identification, the unit root test by augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test proved that the data of error is non-stationary. Then, the first differencing (d=1) was applied and generate 1st order difference of error (Δe) time series. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) plot of the sample data are the basic tools to identify the order of autoregressive (AR) process, p, and the order of moving average (MA) process, q.Then, the ARIMA(p, d, q) is expressed as,

$$\begin{aligned}\Delta e_t &= \delta + \varphi_1(\Delta e_{t-1}) + \varphi_2(\Delta e_{t-2}) + ... + \varphi_p(\Delta e_{t-p}) \\ &+ \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}\end{aligned}, \quad (3)$$

where $\Delta e_t \equiv e_t - e_{t-1}$, δ is the constant, φ_1, φ_2, …, φ_p are the AR parameters, ε_t is the randomly error at time t and $\sim N(0,\sigma^2)$, and θ_1, θ_2, …, θ_q are the MA parameters.

It is found that 1-time lag of PACF was more than the critical value that implies the p and q are basically identified as 1 and 0, respectively, therefore the tentative model was ARIMA(1,1,0). After the identification by numerical estimation, φ_1 equals -0.4319 at a 99% confidence interval level of statistical test. An ARIMA model is not sufficient if there are still linear correlations remain in the residuals [16]. Diagnostic checking by Box-Pierce Chi-Square test verified that ARIMA(1,1,0) is sufficient since no correlation of the residuals. Then, the forecasting PM-10 model at time t, $PM_t^{forecast}$, of the hMLPNN(1,1,1)-ARIMA(1,1, 0) is expressed as,

$$\begin{aligned}PM_t^{forecast} = \Bigg(&\frac{2 \times 0.59}{1+\exp(3.88 \times PM_{t-1}^{Actual} + 1.64)} \\ &- 0.83 \Bigg) + e_{t-1} - 0.4319\Delta(e_{t-1}) + \varepsilon_t\end{aligned}. \quad (4)$$

2.1.2 The hRBFNN-ARIMA model

In this hybrid model, MLPNN is replaced by RBFNN to estimate N_t in the first stage. RBFNN has a structure similar to MLPNN except for using Gaussian function as the radial basis activation function, a_i, in hidden nodes. The input variable of RBFNN is the historical PM-10 data of several days. The number of input nodes corresponding to the dimension of input vectors is determined and selected through the optimization. The number of pattern learning of the observation data (PM_t) of n samples which are fed into P input nodes and one output node of RBFNN(P,Q, 1) generates n–P types i.e., the 1st pattern consists of input [PM_1, …, PM_P] and the output is N_{P+1}, the 2nd pattern consists of input [PM_2, …, PM_{P+1}] and the output is N_{P+2} and the last pattern consists of input [PM_{t-P}, …, PM_{t-1}] and the output is N_t which is expressed as,

$$N_t = \sum_{i=1}^{Q} w_i a_i + \beta_0, \quad (5)$$

where

$$a_i = \exp\left(-\sum_{j=1}^{r} \left(PM_{t-j} - PM_{t-j}^{approx} \right)^2 \Big/ \alpha_i^2 \right), \quad (6)$$

PM_{t-j} is the input for j^{th} node of an input layer, PM_{t-j}^{approx} is the preceding forecast value of j^{th} input which is used as centre of Gaussian function, α_i is the spread parameter of i^{th} node in hidden layer, w_i is the weight between i^{th} node in the hidden layer and the output layer, and β_0 is the bias of the output node.

To avoid the over fitting problem, the train and test samples ratio is also first specified. The designed parameters indicated through RMSE criterion showed that the train: test ratio, the number of input and hidden nodes are 80:20, 1, and 2, respectively. Then, RBFNN was represented by RBFNN(1,2,1) . Later, the unsupervised learning process as K-mean algorithm [17] solved for the centre and variance of Gaussian function. Finally, the gradient descent algorithm, the supervised learning method, adjusted the weights between hidden and output layer. The RBFNN(1,2,1) is mathematical expressed as,

$$\begin{aligned}N_t &= -1.5 \times \exp-[(-0.32 \times PM_{t-1} + 0.83 \\ &- PM_{t-1}^{approx})^2] + 0.83 \times \exp-[(-0.32 \times . \\ &PM_{t-1} + 0.83 - PM_{t-1}^{approx})^2] - 0.032\end{aligned} \quad (7)$$

The residuals from (7) were fed into ARIMA model in the second stage which was constructed similar to the previous Sub-Section 2.1.1. It was found that the ARIMA(0,1,2) is satisfied the statistical test then the *h*RBRNN(1, 2,1)-ARIMA(0,1,2) is hold and expressed as,

$$\begin{aligned} PM_t^{forecast} &= -1.5 \times \exp-[(-0.32 \times PM_{t-1} + 0.83 \\ &- PM_{t-1}^{Approx})^2] + 0.83 \times \exp-[(-0.32 \times \\ &PM_{t-1} + 0.83 - PM_{t-1}^{Approx})^2] - 0.032 \\ &+ e_{t-1} - 0.17 \times (e_{t-2}^{Actual} - e_{t-2}^{Approx}) + \varepsilon_t \end{aligned} . \quad (8)$$

2.2 The hNN-ARIMAX model

The *h*NN-ARIMAX model is mainly composed of two stages (Fig.3). In the first stage, NN operates by using *P*-time lag of historical PM-10 and significant exogenous variables. The generated residuals together with the selected exogenous variables were provided as the input of ARIMAX model in second stage. Two different types of this hybrid model i.e., *h*MLPNN-ARIMAX and *h*RBFNN-ARIMAX were detailed in the next Sub-section 2.2.1 and 2.2.2, respectively.

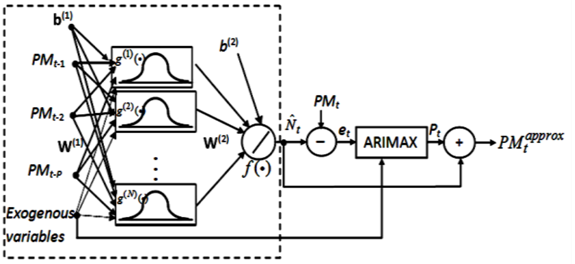

Fig. 3 The PM-10 forecast based on *h*NN-ARIMAX model.

2.2.1 The hMLPNN-ARIMAX model

In this hybrid model, it was resulted that 3-time lag of historical PM-10 and all exogenous variables except for *T* variable are significant correspond to 1 hidden node. Then, the optimized MLPNN is represented by MLPNN ((***X***[7],3),1,1), where ***X***[7]=[*CO*, *O*$_3$, *SO*$_2$, *NO*$_2$, *G*, *W*, *P*, *H*]^T. The residuals from the MLPNN model were fed into ARIMAX model in the second stage. The procedure of ARIMAX model is similar to that of ARIMA model. Let the *L*-time lag of the independent *K*-input variable sequences expressed by {*X*$_1$}, …, {*X*$_K$}, and the first difference of residual (Δ*e*) from (3), the ARIMA(*p*,*d*,*q*)X(*K*) is expressed as follows,

$$\Delta e_t = \delta + \sum_{j=1}^{K}\sum_{i=1}^{L} \mu_{j,i} X_{j,t-i} + \sum_{i=1}^{P} \varphi_i \Delta e_{t-i} + \sum_{i=0}^{Q} \theta_i \varepsilon_{t-i}, \quad (9)$$

where μ_i, φ_i, and θ_i denote the coefficient parameters, *K*, *P*, and *Q* denote the maximum time lag related to the independence sequences, residuals, and the randomly error, respectively, and δ is a constant. From the optimization, it is found that *L* equals 1 is sufficient. Following by the ARIMA procedure in Section 2.1, ARIMA (1,1,0)*X*[7] was satisfied through the statistical tests. The hybrid model was then resulted by *h*MLPNN((***X***[7],3),1)-ARIMA(1,1,0)***X***[7] which is expressed as

$$\begin{aligned} PM_t^{forecast} = \Big(&\frac{2 \times 0.4354}{1 + \exp((-2\mathbf{W}_1 \times \mathbf{X} - 1.64))} \\ &- 0.4354 \Big) - 0.318 + e_{t-1} - 0.407 , \\ &\times \Delta(e)_{t-1} + \mathbf{W}_2 \times \mathbf{X} + \varepsilon_t \end{aligned} \quad (10)$$

where

$$\begin{aligned} \mathbf{W}_1 \times \mathbf{X} = 1.30 \times CO_{t-1} &+ 1.09 \times O_{3,t-1} + 0.29 \\ \times SO_{2,t-1} &- 1.96 \times NO_{2,t-1} + 0.44 \quad , \\ \times GW_{t-1} &- 0.48 \times P_{t-1} - 1.89 \times H_{t-1} \end{aligned} \quad (11)$$

and

$$\begin{aligned} \mathbf{W}_2 \times \mathbf{X} = 0.22 \times CO_{t-1} &- 0.07 \times O_{3,t-1} - 0.01 \\ \times SO_{2,t-1} &+ 0.95 \times NO_{2,t-1} + 0.15 \quad , \\ \times GW_{t-1} &- 0.03 \times P_{t-1} + 0.52 \times H_{t-1} \end{aligned} \quad (12)$$

2.2.2 The hRBFNN-ARIMAX model

In this hybrid model, it was resulted that 3-time lag of historical PM-10 and all exogenous variables except for *T* were significances correspond to 10-selected hidden node. Then, RBFNN model is represented by RBFNN((*X*[7], 3),10,1). The residuals from the RBFNN together with *X*[7] were continuously fed into ARIMAX model in the second stage. AR(0)I(1) MA(2)*X*[7] was satisfied from the statistical tests. Finally, the hybrid model, *h*RBFNN((X[7] ,3),10,1)-ARIMA(0,1,2)X[7] is expressed as follow,

$$\begin{aligned} PM_t^{forecast} = \sum_{j=1}^{8} \Bigg(&w_{jo} \times \exp- \Bigg(\sum_{i=1}^{7} w_{ij} X_i + \\ &\sum_{k=1}^{3} w_{kj} PM_{t-k} + b_j - c_j \Bigg)^2 \Bigg) + b_o \quad , \\ &+ e_{t-1} - 0.1968 \times \left(e_{t-2}^{Actual} - e_{t-2}^{Approx} \right) \\ &+ \mathbf{W}_3 \times \mathbf{X} + \varepsilon_t \end{aligned} \quad (13)$$

where a number of w_{ij} and w_{kj} parameters from this model are not shown here, and

$$\begin{aligned}\mathbf{W}_3 \times \mathbf{X} = {} & 1.30 \times CO_{t-1} + 1.09 \times O_{3,t-1} \\ & + 0.29 \times SO_{2,t-1} - 1.96 \times NO_{2,t-1} \\ & + 0.44 \times GW_{t-1} - 0.48 \times P_{t-1} - 1.89 \times H_{t-1}\end{aligned} \quad . \quad (14)$$

The forecast results of the proposed hybrid forecast models mentioned above were shown in the next Section with the discussions.

3. Results and discussions

The hybrid forecast models formulated in Section 2 are validated with the testing data in 2015 in order to indicate how well the model has been trained. The forecast results of the **Model (1)** *h*MLPNN(1,1,1)-ARIMA(1,1,0), **Model (2)** *h*RBFNN(1,2,1)-ARIMA(0,1,2), **Model (3)** *h*M LPNN((*X*[7],3),1)-ARIMA(1,1,0)*X*[7] and **Model (4)** *h*RBFNN((*X*[7],3,1),10)-ARIMA(0,1,2)*X*[7] are illustrated in Fig. 4-7, respectively. It is indicated that the hybrid forecast **Model** 1-4 can estimate the high PM-10 very well. Furthermore, it is seen that **Model (4)** outperforms **Model (1)**, **Model (3)** and **Model (2)**, respectively.

By using testing data in 2016, the proposed hybrid forecast models were compared to the previous models referenced in [12] including ARIMA(4,1,3), ARIMA(4,1,3)*X*[1], MLPNN(1, 1,1), RBFNN(1,2,1), *h*ARIMA(4,1,3)-MLPNN ([2,1],1,1), ARIMA(4,1,3)-RBFNN([1,2],2,1), *h*RBFNN(1,2,1)-ARIMA(0,1,2) and *h*ARIMA(4 ,1,3)*X*(1)-RBFNN([1,3],2,1). The performance of models is assessed using MAE, RMSE, AIC and the number of model parameters (Table 2). It is found that the *h*NN-ARIMA/ARIMAX model gives the best forecast at the high PM-10 level as well as the overall average whereas the *h*ARIMA/ARIMAX-NN model gives the worst forecast. It indicates that the prior processing either linear or nonlinear is the significant issue. Since main pattern of the PM-10 problem is nonlinear and complex then the NN as the nonlinear model should firstly perform and continue with the ARIMA/ARIMAX model as the linear model. However, in general, there is no any theoretical guarantee which model is better because it depends on a particular problem. For ranking the PM-10 forecast models with AIC [18] that seeks a model which has a good fit but few parameters, it is defined as,

$$AIC = n\log\left(\frac{\sum_{n} e_i^2}{n}\right) + 2K, \qquad (15)$$

where e_i is the normally distributed residuals, n is the number of test data, and K is the total number of parameters in the model. The preferred model is the one with the minimum AIC.

Although **Model (4)** provided the most accurate forecast considered from the RMSE and MAE. However, a number of its parameters caused complexity and high computation time is not suitable and reliable in the real practice. From the model assessment through AIC, it was shown that **Model (1)**, **Model (3)** and **Model (2)** are considered as the best model whereas **Model (4)** is the worst model (Table 2). To forecast 5-day PM-10 ($PM10_{t+1}$, $PM10_{t+2}$,…, and $PM10_{t+5}$) by the **Model (1)**, hybrid MLPNN-ARIMA model, the results showed that it gives accurate and reliable for only 1-day forecast while the forecast trend increasingly saturates for 2-day to 5-day forecast.However it was seen from Table 1 that most of the variables have a wide range of relative change (ratio of variance compared to range) from 0.06 (of SO_2) to 4.09 (of PM-10). The proposed hybrid forecast model based on one rule describing the dynamic change of the input variables probably would not be sufficient to provide the most accurate forecast. To complete our work, the forecast model based on an ANFIS model was further implemented to compare the performance against the *h*MLPNN-ARIMA model as the best model and others.

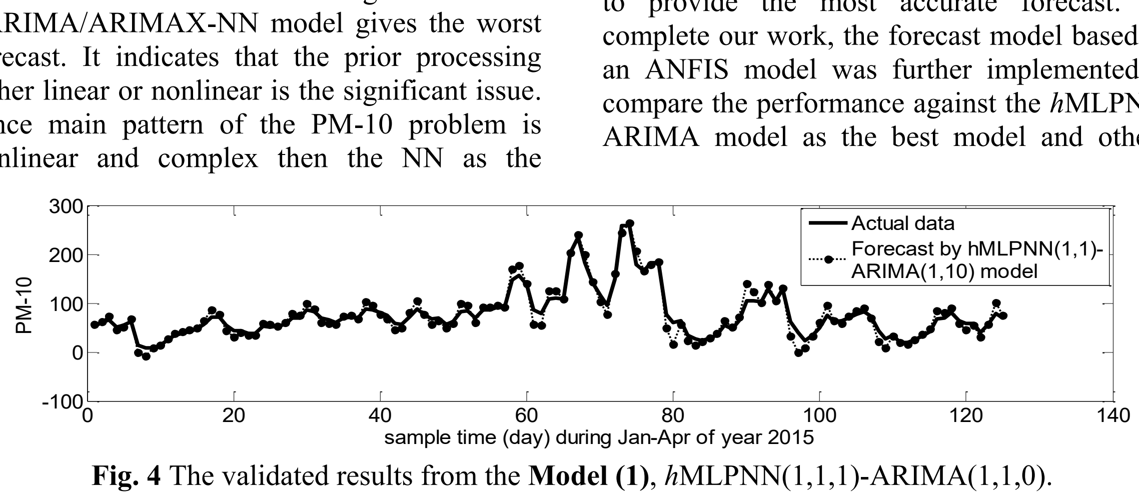

Fig. 4 The validated results from the **Model (1)**, *h*MLPNN(1,1,1)-ARIMA(1,1,0).

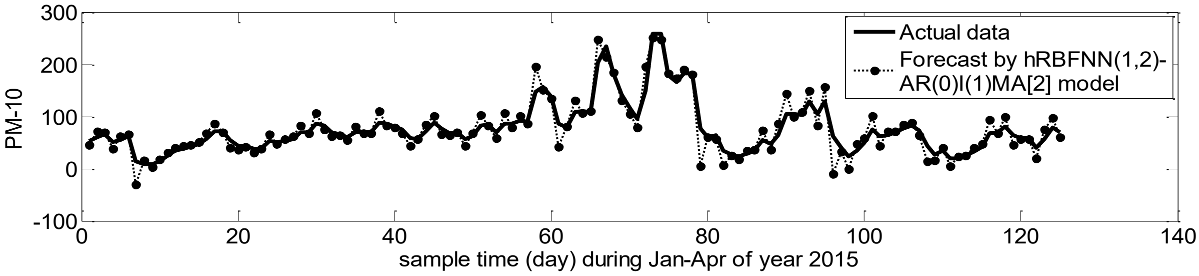

Fig. 5 The validated results from the **Model (2)**, *h*RBFNN(1,2,1)-ARIMA(0,1,2).

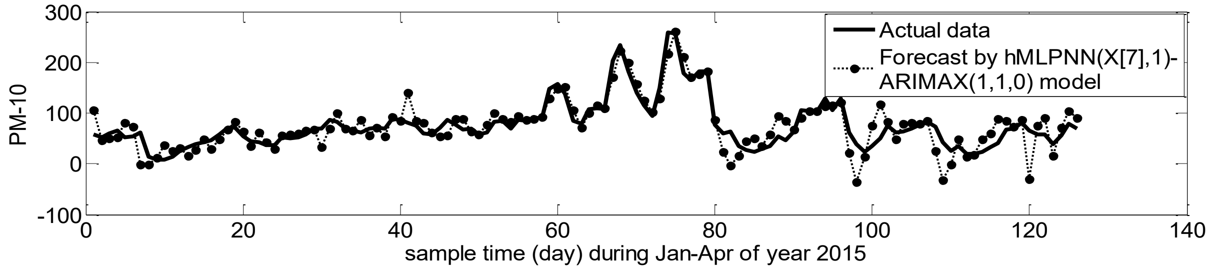

Fig. 6 The validated results from the **Model (3),** *h*MLPNN((***X***[7],3),1)-ARIMA(1,1,0)***X***[7].

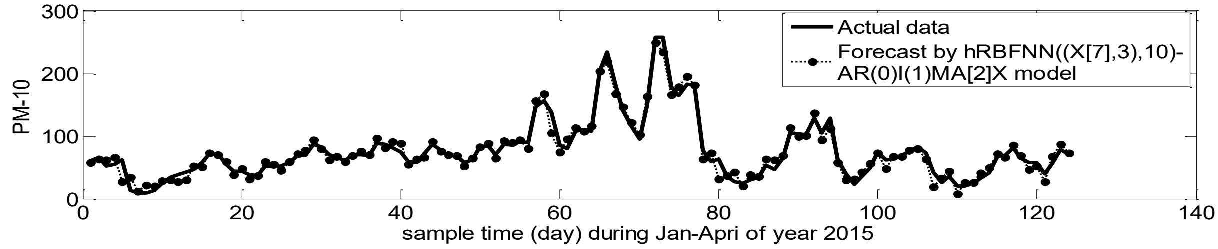

Fig. 7 The validated results from the **Model (4),** *h*RBFNN((***X***[7],3),10)-ARIMA(0,1,2)***X***[7].

Table 2 Performance comparison of the proposed PM-10 forecast models with existing models.

Model Ranking	RMSE	MAE	AIC	The number of model parameters
1) **Model (1)***h*MLPNN(1,1,1)-ARIMA(1,1,0)	12.2	8.9	19.8	5 (*w* =2, *b*=2, AR=1)
2) **Model(3)***h*MLPNN((X[7],3),1,1)-ARIMAX(1,1,0)	13.8	9.3	50.1	14 (*w*=11, *b*=2, AR=1)
3) **Model(2)***h*RBFNN(1,2,1)-ARIMA(0,1,2)	17.3	11.7	62.1	8 (*w*=4, *b*=3, MA[2]=1)
4) ARIMA(4,1,3)X(1)	23.1	15.7	91.8	8 (AR=4, MA=3, *X*=1)
5) MLPNN(1,1,1)	25.4	16.0	93.6	4 (*w*=2, *b*=2)
6) ARIMA(4,1,3)	24.7	16.6	96.9	7 (AR=4, MA=3)
7) RBFNN(1,2,1)	25.7	16.0	100.8	7 (*w*=4, *b*=3)
8) *h*ARIMA(4,1,3)*X*(1)-MLPNN([1,1],1,1)	24.6	16.0	108.6	13 (8,*w*=3,*b*=2)
9) *h*ARIMA(4,1,3)-MLPNN([2,1],1,1)	25.5	15.3	112.2	13 (7,*w*=4,*b*=2)
10) *h*ARIMA(4,1,3)*X*(1)-RBFNN([1,3],2,1)	26.4	15.0	131.7	21(8,*w*=10,*b*=3)
11) *h*ARIMA(4,1,3)-RBFNN([1,2],2,1)	28.2	17.9	132.7	18 (7,*w*=8,*b*=3)
12) **Model(4)***h*RBFNN((*X*[7],3),10,1)-ARIMA*X*(0,1,2)	9.8	7.4	230.6	122(*w*=110,*b*=11,1)

The ANFIS structure is obtained by embedding the fuzzy inference rules into the framework of adaptive networks [19]. For 5-day PM-10 forecast i.e., PM_{t+1}, …, and PM_{t+5}, five different ANFIS models used the historical *P*-time lag PM-10 (PM_{t-1}, …, and PM_{t-P}) and exogenous variables, $\boldsymbol{X}$ = [CO, O_3, SO_2, NO_2, GW, P, T, H]^T as the input variables and the number of membership functions (MFs) is $N_{PM(t-1)}$,…, $N_{PM(t-P)}$, N_{CO}, …, and N_H, respectively. It gives the output as the PM_{t+i} where i =0, …, 4. The ANFIS model typically includes 5 layers in which nodes of the same layer have similar MF

(Fig. 8). The stepwise procedures are detailed as follows,

- Layer 1: The input variables are normalized to the range [0, 1].
- Layer 2: The adaptive nodes consisted of Gaussian MFs (GMF) transforms the crisp of input variable to the degree of MF (μ) of fuzzy set PM_{t-1} = {*NB*, *NS*, …, *PS*, *PB*} ,…, *H* = {*NB*, *NS*, …, *PS*, *PB*} which are denoted by $\mu_{PM(t-1)_x},\ldots,\mu_{H_x}$, respectively

  where *x* is the fuzzy variable, $x \in$ {*NB*, *NS*, …, *PS*, *P*B} and NB, NS, PS and PB are referred as negative big, negative small, positive small and positive big, respectively.
- Layer 3: The number of constructed IF-THEN rules, N_{rule}, is $N_{PM(t-1)} \times \ldots \times N_{PM(t-P)} \times N_{CO} \times \ldots \times N_H$. For the first-order Sugeno fuzzy, the fuzzy rule is shown for example below:

  Rule i: IFPM_{t-1} is *NB* and … and PM_{t-P} is NB and ,…, and *H* is *NB*,

  THEN $\theta_i = p_1(PM_{t-1}) + \ldots + p_P(\mathrm{PM}_{t-P}) + \ldots + p_H(H) + p_0$.

  The number of system parameters is $(N_{PM(t-1)} + \ldots + N_{PM(t-P)} + N_{CO} + \ldots + N_H) \times (2 + N_{rule})$ including GMFs parameters (centre, *c* and variance, σ^2) and consequence parameters ($p_1,\ldots, p_P,\ldots, p_H$, and p_0).
- Layer 4: For the rule premise evaluation, the product for *T*-norm (logical *and*) is chosen and resulted the weight values as

  $$w_j = \mu_{PM(t-1)_X} \times \ldots \times \mu_{PM(t-P)_X} \times \ldots \times \mu_{H_X}$$
  (16)

  where $j = 1, 2, \ldots, N_{rule}$.

  The consequence rules corresponding with the weight values are posed to the layer 5 for the implication evaluating.
- Layer 5: The forecast output is averagely calculated by

$$PM_{t+i} = \sum_{j=1}^{N_{rule}} w_j \theta_j \Big/ \sum_{j=1}^{N_{rule}} w_j \ . \qquad (17)$$

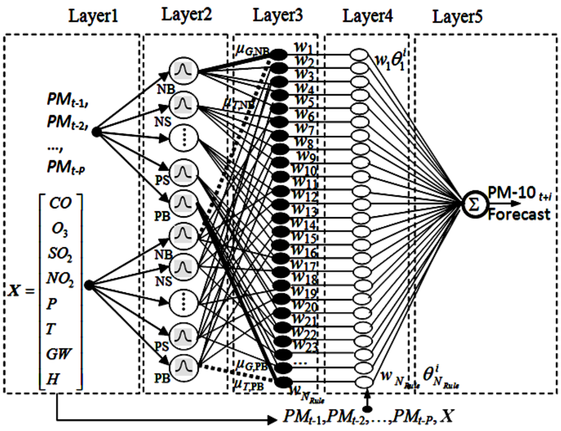

Fig. 8 The PM-10 forecast based on ANFIS model.

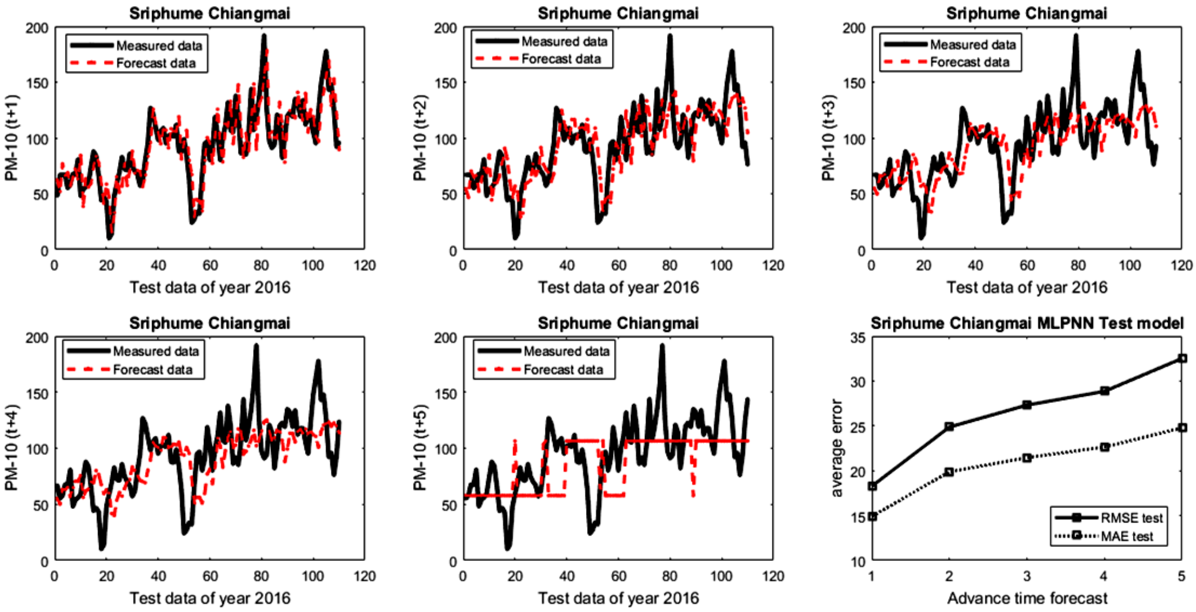

Fig. 9 A 5-day PM-10 forecast by using the optimal *h*MLPNN-ARIMA model.

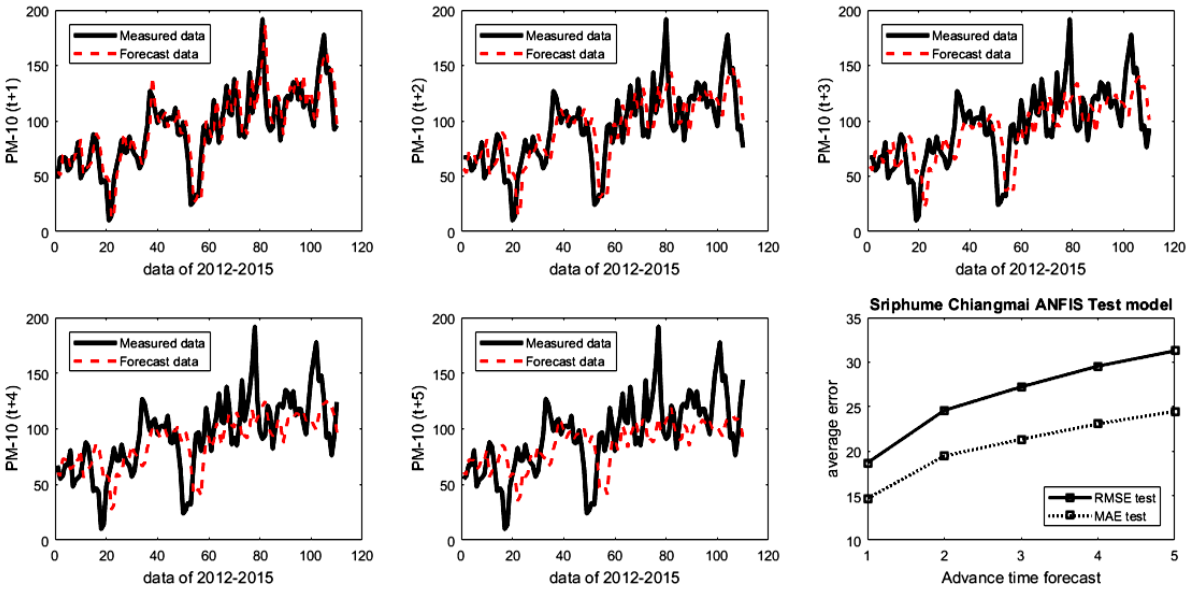

Fig. 10 A 5-day PM-10 forecast using the ANFIS model.

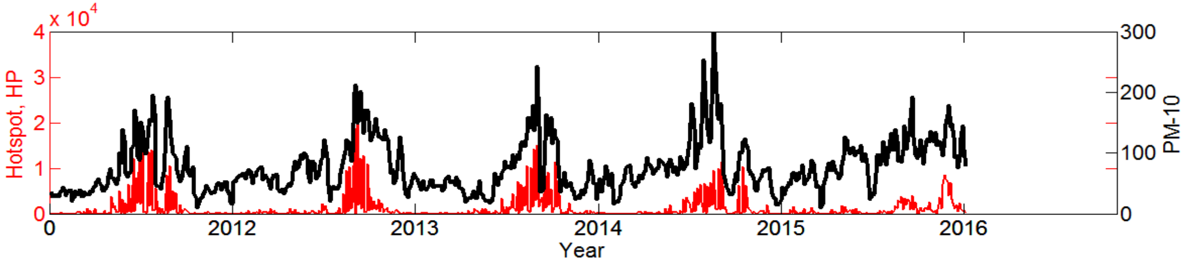

Fig. 11 The correlation between hotspots and PM-10 in the period January-April during 2012-2016 of Chiang Mai Province, Thailand.

The outputs of the adaptive nodes depend on the adjusted parameters which should be iteratively adjusted to minimize the error through the learning rule. For evaluation by ANFIS editor in MATLAB program, a large number of parameters and also the MFs can lead to very slow convergence or terminated program. To optimize the ANFIS through the experimental design, it was found that the significant variables for 1-day forecast is $\{PM_{t\text{-}1}$ and $H\}$ and for 2-5 days forecast are $\{PM_{t\text{-}1}, T\}$.

To construct the ANFIS model, the number of MFs of $PM_{t\text{-}1}$, H and T were set to 5 corresponds to 25 fuzzy rules. The number of system parameters according to 25 rules yields 2×5+2×5+3×5×5 = 95 parameters which were adjusted through a hybrid learning rule combining the BP gradient descent and a least-squares method. By using the test data in 2016, the ANFIS gives the forecast performance slightly less than that of the *h*MLPNN-ARIMA (Fig.9-10) and requires more cost computation and a number of parameters. The execution time of forecast **Model** 1-4 is less than a second since they have exactly closed-form expressions referred to (2), (8), (10) and (13), respectively. However, the different complexity in those models, **Model** 4 takes more execution time than **Model** (3), (2) and (1), respectively. While the ANFIS model related to the fuzzy IF-THEN rules transformations and has not exactly closed-form expression requires more running time of program 3-4 times that of the hybrid **Model** 1-4.

For the application, the optimal hybrid PM-10 forecast model was online embedded through some application software beneficial for the people in Chiang Mai and nearby to instantaneously get the information and prepare in advance.

4. Conclusions

This work presents the various methodologies to formulate the PM-10 forecast model during high season in Chiang Mai (Thailand). The single models including, ARIMA, ARIMAX, MLPNN and RBFNN and the hybrid models including, *h*ARIMA(X) - MLPNN/ RBFNN and *h*MLPNN/ RBFNN-ARI MA(X) were implemented. The significant historical PM-10 data, related meteorological data and correlated toxic gases were taken as the input variable. The performance of the models in terms of accuracy identified by RMSE and MAE is indicated that *h*RBFNN-ARIMAX model performed better than the rest. However, trade-off between accuracy and reliability through AIC, it indicated that the *h*MLPNN-ARIMA model was the optimal model whereas the *h*MLPNN-ARIMAX and the *h*RBFNN-ARIMA model were the sub-optimal models. The *h*RBFNN-ARIMAX model was identified as the worst model. To further verify the forecast performance between the optimal *h*MLPNN-ARIMA model and the ANFIS based on PM-10 forecast model, it was supported through AIC that the *h*MLPNN-ARIMA model gives slightly more accurate than that of the ANFIS model but more reliable.

In the future, one possible extension of this research is to improve the forecast model by including hotspot as the other exogenous variable since it was demonstrated the great impacts of the burning in this area which is evident through the correlation between the hotspots, counted from Terra and Aqua MODIS (Moderate Resolution Imaging Spectroradiometer) and PM-10 level (Fig. 11). In addition, the forecast is extended for the other upper provinces in Northern Thailand where PM-10 produce great impacts on human health.

Acknowledgements

The author directly acknowledges the research fund support from the research institute of the North-Chiangmai University. In addition, we are grateful to the sub-division of Thai Meteorological Department and PCD for the collected valuable statistical data.

References

[1] Chaloulakou, A., Grivas, G., and Spyrellis, N. Neural network and multiple regression models for PM10 prediction in Athens: A comparative assessment. *Journal of the Air & Waste Management Association*, 2003; 53:1183–1190.

[2] Diaz-Robles, L.A., Ortega,J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, G.J., and Moncada-Herrera, J.A. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, 2008; 42: 8331–8340.

[3] Goyal, P., Chan, A.T. and Jaiswal, N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment*, 2006; 40: 2068–2077.

[4] Liu, P.W.G. Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis. *Journal of the American Statistical Association*, 2009; 92(439): 2104-2113.

[5] Sotomayor-Olmedo, A., Aceves-Fernández, M.A., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arreguín, C.J.M., and Vargas-Soto, J.E. Forecast urban air pollution in Mexico City by using support vector machines: A kernel performance approach. *International Journal of Intelligence Science*, 2013; 3: 126–135.

[6] Chuentawat, R., Kerdprasop, N., and Kerdprasop, K. The forecast of PM10 pollutant by using a hybrid model. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 2017; 6(3):128-132.

[7] Kanabkaew, T. Prediction of hourly particulate matter concentrations in Chiangmai, Thailand using MODIS aerosol optical dept and ground-based meteorological data. *Environment Asia*, 2013; 6(2): 65-70.

[8] Amphanthong, P. and Busababodhin, P. Forecasting PM-10 in the upper northern area of Thailand with Grey system theory. *Burapha Journal*, 2015; 20(1): 65-70.

[9] Wongsathan, R., and Seedadan, I. Prediction modeling of PM-10 in Chiangmai city moat by using artificial networks. *Journal of Applied Mathematics and Mechanics*, 2015; 781: 628-631.

[10] Wongsathan, R. and Seedadan, I. A hybrid arima and neural networks model for PM-10 pollution estimation: The case of Chiang Mai city moat area. *Procedia Computer Science*, 2016; 86: 273-276.

[11] Wongsathan, R., Seedadan, I. and Wanasri, S. Hybrid forecast model for PM-10 prediction: A case study of Chiang Mai city of Thailand during high season. *KKU Engineering Journal*, 2016; 43(S2): 203-206.

[12] Wongsathan, R. and Chankham, S. Improvement on PM-10 forecast by using ARIMAX and hybrid ARIMAX and neural networks model for the summer season in Chiang Mai. *Procedia Computer Science*, 2016; 86: 277-280.

[13] Wongsathan, R. Performance of hybrid fuzzy and neuro-fuzzy controller with mohga based mppt for a solar pv module on various weather conditions. *Engineering Journal Chiang Mai University*, 2017; 24(2): 161-175.

[14] Prasad, K., Gorai, A.K., and Goyal, P. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmospheric Environment*, 2016; 128: 246-262.

[15] Oprea, M., Mihalache, S.F., and Popescu, M. Computational intelligence-based PM2.5 air pollution forecasting. *International Journal of Computers Communications & Control*, 2017; 12(3): 365-380.

[16] Wang, X. A hybrid neural network and arima model for energy consumption forecasting. *Journal of computers*, 2012; 7(5): 1184-1190.

[17] Sing, J.K., Basu, D.K., Nasipuri, M., and Kundu, M. Improved K-means algorithm in the design of RBF neural networks. IEEE TENCON 2003: Conference on Convergent Technologies for the Asia-Pacific Region, October 15-17, 2003, Bangalore, India, 58-70.

[18] Burnham, K.P., and Anderson, D.R. Model Selection and Multimodel Inference. New York: Springer-Verlag, USA, 2002.

[19] Takagi, T., and Sugeno, M. Fuzzy identification of system and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1985; 15(1): 116.