EnvironmentAsia
The international journal by the Thai Society of Higher Education Institutes on Environment

EnvironmentAsia 13(2) 2020 15-30
DOI 10.14456/ea.2020.27
ISSN 1906-1714; ONLINE ISSN: 2586-8861

Application of Daily Air Pollutant Index Forecasting Model Based on Semi-Empirical Statistical Theory: Case Study in Hanoi, Vietnam

Pham Thi Thu Ha[1], Pham Thi Viet Anh[1], Do Manh Dung[2], Duong Ngoc Bach[3], Phan Thu Trang[3], Nguyen The Hung[3], and Pham Ngoc Ho[3*]

[1] *Faculty of Environmental Sciences, University of Science, Vietnam National University, Hanoi*
[2] *VINACOMIN Informatics, Technology, Environment Joint Stock Company, Unit Resources A, B15, Dai Kim Ward, Hoang Mai district, Hanoi, Vietnam*
[3] *Research Center of Environmental Monitoring and Modeling, University of Science, Vietnam National University, Hanoi*

*Corresponding author: hopn2008@yahoo.com.vn
Received: November 26, 2019; Revised: December 12, 2019; Accepted: February 9, 2020

Abstract

This article illustrates Pham Ngoc Ho's forecasting model applied for daily air pollutant index, including PM_{10}, CO, NO_2 and O_3. The authors have tested the model by analyzing 24-hour-per-day continuous monitoring data in 2017 – 2018 from the Nguyen Van Cu permanent monitoring station to forecast daily air pollution index in Hanoi. The results have demonstrated that our model forecasts the air pollution index with the efficiency of 75-95% and 85-98% for the case of respectively 1-hour, 8-hour and 24-hour average in a day. In comparison with this model, the following models are applied and cited: Hanna SR's simple statistical model which is tested using the same monitoring data from Nguyen Van Cu permanent monitoring station, the interpolation/extrapolation model of the author Duong Ngoc Bach which used monitoring data in 2012 at Nguyen Van Cu permanent monitoring station and Pongpiachan and Paowa's model of pollutants interacting with meteorological factors applied in Chiang-Mai, Thailand in 2015. The results of these 3 models have shown that, in contrast of Hanna SR's model that has relatively low accuracy, the remaining models have high accuracy. However, the model we use has an outstanding advantage of forecasting the air pollution index according to the daily forecast of meteorological factors on the mass media. This is a new approach that has never been reportedly applied to any contemporary modelling. This is the goal of the study.

Keywords: Modeling; Daily air pollution index; Model's efficiency

1. Introduction

Recently in Vietnam, there have been some studies applying small-scale forecasting model for many urban areas: Berliand's diffusion model K, applied for point and non-point sources (Berliand, 1985); Gauss's diffusion model, applied for point and non-point sources (Pasquill, 1974; Schonoor, 1996); Statistical model based on theory of random functions, using 24-hour-per-day continuous monitoring data to forecast in real-time (Kazakevits, 1971; Pham Ngoc Ho, 2016); Simple Statistical Model estimating concentration C(t) from initial data (Hanna, 1982).

The models of Berliand and Gauss can mainly be applied to simulate pollutants transport processes from point and non-point sources in a given space at certain times. Their accuracy, however, relies heavily on each pollutant's emission factor, which is still very difficult to be identified in Vietnam, and surrounding climatic condition, which varies from country to country. Besides, there have been studies on the effects of polycyclic aromatic hydrocarbons (PAHs), PM_{10}, $PM_{2.5}$ (Pongpiachan *et al*., 2015, 2017) and the relationship of meteorological factors and pollution parameters using

advanced statistical models, including t-tests, Analysis of Variance (ANOVA), Multiple Linear Regression Analysis (MLRA) and Incremental Lifetime Particulate Matter Exposure (ILPE) (Pongpiachan, 2014, Pongpiachan and Paowa, 2015). Those research results are of high practical value and are a useful reference for experts in forecasting the quality of environmental parameters by mathematical-chemical model. Many interpolation or extrapolation models are usually used to repair some raw data absence (Duong Ngoc Bach, 2012, 2016; Tran Thi Thu Huong, 2017). Forecasting model of air pollutants index based on semi-empirical statistical theory for SO_2, NO_2, CO, O_3 parameters and dust (PM_{10} and $PM_{2.5}$) according to meteorological factors (pressure, temperature, humidity, wind speed) and average daily concentration of each 24-hour automatic monitoring parameter at different monitoring sites (Pham Ngoc Ho, 2017).

The forecasting models of Berliand and Gauss need to have emission factors from point/non-point sources for each pollutant; thereby they cannot be applied under Vietnam circumstance. Therefore, here we use only 4 models that can be applied in practice to calculate and compare, specifically, Pham Ngoc Ho's model, Hanna SR's model, the interpolation/extrapolation model repair data absence and the model of pollutants, chemicals (SO_2, NO_2, CO and O_3) and dust (PM_{10}, $PM_{2.5}$) interacting with meteorological factors applied by Pongpiachan and Paowa in 2015.

2. Methodology

2.1 Daily air pollutant index forecasting model based on semi-empirical statistical theory

2.1.1 Materials

The automatically monitored data for 12 consecutive months in 2017 – 2018 at Nguyen Van Cu monitoring station were used to calculate the daily air pollution index q. The monitoring equipment is from HORIBA-Japan. Different model numbers are used to monitor different parameters, which are listed as follows:

The station is located at coordinates 21º02'56.3"N, 105º52'58.8"E, and 2.5m from the ground; its detector is 3-3.5m above the ground. The width of pavement is 5m (adjacent to the highway).

The station has been in operation since 2010, monitoring the following parameters: meteorological factors (wind, wind direction, pressure, temperature, humidity, and solar radiation), basic parameters (NO, NO_2, NO_x, SO_2, CO, O_3, PM_{10}, $PM_{2.5}$ and PM_1). In addition, the station has a standardized system and devices for data collection, reception and transmission. The station is managed by Center for Environmental Monitoring under Vietnam Environment Administration and is allocated budget for operation maintenance to ensure the accuracy of the data. The principles of operation of the device, measurement methods, chemical reactions and operation of each module for gas and dust are described in detail in manufacturers' brochures (HORIBA Process and Environment; GRIMM Aerosol Technik).

Criteria for parameters selection and homogeneous process of data:

The selected parameters must be in Vietnam National Regulation on Ambient Air Quality (QCVN) and the series of monitoring data must satisfy at least 70% of days in a year that have continuous data. To homogenize data, it is necessary to: (1) Eliminate data anomalies contrary to natural laws, (2) Compare the data of the same monitoring station and environmental data with meteorological data, (3) Eliminate invalid data including data in the time of adjusting the device, data with negative values, data with continuously equal values, etc.

Based on the reasons mentioned above, the automatically monitored data of Nguyen Van Cu station in 2017-2018 attributes to only 4 parameters: NO_2, CO, O_3 and dust PM_{10} ($\mu g/m^3$) that meet the data selection criteria, so they have been selected to calculate the input of the forecasting model. The process of inputting data to calculate average concentration value of each parameter (1- hour average, 8- hour average and 24- hour average) is conducted according to the national standard/regulation of Vietnam (MONRE 2013). As for average hourly data, the highest concentration value

of each parameter monitored during 24 hours of a day was used to calculate the index q(1h); As for 8-hour average data, the average values of 8 observations (there are three 8-hour average values per day) were taken, then the highest value from these 8-hour average values was used to calculate the index q (8h); as for 24-hour average data, the average value of 24 observations was taken to calculate the index q (24hr).

2.1.2. Pham Ngoc Ho's daily forecasting method for air pollutant index based on semi-empirical statistical theory (Pham Ngoc Ho, 2017)

Approach

Semi-empirical statistical theory is a theory based on dimensions to standardize input parameters of a model (such as wind speed, air pressure, humidity, temperature and pollutants' concentrations depends on time t), combining with applying statistical theory to analyze 24-hour monitoring data as the input and to correct the model.

Standardization has been proceeded with a purpose that the left and right sides of the equation will be similarly dimensional (having same units) or non-dimensional like the air pollution or air quality index.

Scientific basis of the model

Firstly, we use the general gas equation (Pham Ngoc Ho *et al.,* 2011)

$$P = \rho RT \qquad (1)$$

where: P is air pressure (mb),

ρ is density of the air layer right above the ground,

R is the specific gas constant of dry air, calculated via universal gas constant according to formula

$R = \frac{R^*}{\mu}$, $R^*=8.314\times10^7$ ec/mol.°C

with μ is gram particle density of the surveyed pollutants, therefore R can be seen as being known already.

From formula (1), we can see that the concentration of recorded pollutant (such as PM_{10}, $PM_{2.5}$, SO_2, NO_x, CO, O_3) is in direct ratio with air pressure P and in inverse ratio with daily average temperature T.

Physically, the concentration of pollutant (in average 1 hour) is in inverse ratio with daily average wind speed u (the lower the daily average wind speed u, the higher the pollutant's concentration); and higher humidity means the pollutants could absorb more water vapor, resulting in reducing pollutants' concentration).

On that account, by using semi-empirical theory to standardize input parameters (including air pressure p, wind speed u, temperature T and humidity f and pollutants' concentration changing over time t, air quality index q(t) of surveyed parameters is a non-dimensional quantity (having no unit). Therefore, the model of forecasting-pollution index q(t) is:

$$q(t) = \frac{P}{P_0}\times\frac{f_0}{f}\times\frac{u_0}{u}\times\frac{T_0}{T}\times q_0(\text{avg24h}) \left[a\left(\frac{t}{t_0}\right)^6 + b\left(\frac{t}{t_0}\right)^5 + c\left(\frac{t}{t_0}\right)^4 + d\left(\frac{t}{t_0}\right)^3 + e\left(\frac{t}{t_0}\right)^2 + g\left(\frac{t}{t_0}\right) + h\right] \qquad (2)$$

where:

$t_0=24^h$ is daily period

P_0, f_0, u_0 and T_0, are daily 24-hour average values of contemporary days; P, f, u and T are forecasted values for the following days.

q_0 is identified daily 24-hour average value of contemporary days.

The coefficients a, b, c, d, e, g and h are non-dimensional factors regressed by 6th degree polynomial function. Polynomial function of degree 6 was selected because many latest studies (Duong Ngoc Bach 2016, Tran Thi Thu Huong 2017) have indicated that the distribution graph of 1-hour average concentration of pollutant changing over time is simulated by this function has significantly higher accuracy than others of degree from 1 to 5.

2.1.3. Process of forecasting calculation

Calculating forecasting index q(t) of surveyed pollutants over time t

To calculate forecasting index $q_f(t)$ of January 2nd, 2018, we need input data of January 1st, 2018 and the processed 24-hour-per-day consecutive monitored data of at least 70% of total days in 2017 to regress and calculate, following 3 steps:

Step 1: Regressing 6th degree polynomial function

With the processed data for 2017 we can calculate hourly average value over standardized time: $t_j = \frac{t_i}{24}$, i=1,2,..., 24h by formula:

$$Q(t_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{C(t_j)}{C^*(1h)} \quad (3)$$

where $C(t_j)$ is the monitored value at the standardized time t_j (i=1,2, ..., 24h), n is the number of actual monitoring days in 1 year.

Next, we use equation (3) and Excel software to regress 6th degree polynomial function ($at^6 + bt^5 + ct^4 + dt^3 + et^2 + gt + h$) to identify its coefficients a, b, c, d, e, g, and h.

If correlation coefficient of regression $R^2 \geq 0.75$, the results of identified coefficients could be accepted. In case of $R^2 < 0,75$, data from one year before need to be added to the calculation. The average result of the 2 consecutive years is used to identify regression coefficients to guarantee the statistical stability with $R^2 \to 1$.

By substituting the identified regression function into the original forecasting equation (2), we have forecasting equation with identified coefficients a, b, c, d, e, g, and h.

Step 2: Model Correction

To correct the forecasting equation, we need to calculate the correcting factors.

Supposing we need to forecast for January 2nd, 2018, we use the input data of January 1st, 2018 with respective time t (t=1,2, ..., 24h).

The correction factor α(t) at each time t is calculated by:

$\alpha(t) = q_m(t)$ of January 1st, 2018 – $q_f(t)$ of January 1st, 2018 (4)

where q_m is the monitored value and q_f is the forecasting value calculated by using the forecasting equation.

To calculate the corrected forecast value $q_c^*(t)$for the January 2nd, 2018, we use the equations (21), (22), (23) and (24).

Step 3: Evaluate relative error of corrected model

Since α(t) could be positive, negative or equal to 0, relative error $\varepsilon(t)$ of each time t (t=2, 3, ..., 24h) can be calculated by the following fundamental statistical formula:

$$\varepsilon(t) = \frac{\left|q_m(t) - q_c^*(t)\right|}{q_m(t)} \quad (5)$$

Forecasting 24-hour average pollution index

Take the 24-hour average of pollution index from equation (21) and (22) we have:

$$\overline{q_c^*(t)} = \frac{1}{24}\sum_1^{24} q_c^*(t_i) \quad (6)$$

Multiply both sides of (6) with in which C* (1h) = const (which is the 1- hour average standard of the parameter selected for forecasting), we have:

$$\frac{\overline{q_c^*(t)}}{C^*(24h\ average)} \times C^*(1h) = \frac{1}{C^*(24h\ average)} \times \frac{1}{24}\sum_1^{24}\frac{C(t_i)}{C^*(1h)} \times C^*(1h) = \frac{24h\ average\ of\ C(t_i)}{C^*(24h\ average)} \quad (7)$$

In which, $C(t_i)$ is the forecasted value at the time t_i Formula (7) is used to calculate the value of 24-hour pollution index.

To evaluate the general error of 24-hour average value, we use ε^* calculated by the formula:

$$\varepsilon^* = \frac{1}{24}\sum_{i=1}^{24}\left|\varepsilon(t_i)\right| \quad (8)$$

The regression and calculation processes can be applied for forecasting 8-hour average pollution index (equation (23) and (24)).

2.2 Hanna SR Statistical Model

2.2.1 Materials

The data used is the same as stated in Section 2.1.1.

2.2.2 Scientific basis of the model

Formula: $C_t(x,y,z,t) = C_0(x,y,z,t_0)\left(\frac{t_0}{t}\right)^{\alpha}$ (9)

In which, C_t, C_0 are pollutant concentrations at the permanent automatically monitoring site (x, y, z) over time t and t_0; α is the coefficient calculated from the monitoring data according to the time t of the day.

To determine the empirical coefficient α, proceed to naturally logarithmic two sides of (9):

$$\ln C_t = \ln C_0 + \alpha \ln \frac{t_0}{t} \Leftrightarrow \alpha_t = \frac{\ln \frac{C_t}{C_0}}{\ln \frac{t_0}{t}} \quad (10)$$

Example: Apply to 1 January 2017, calculate αt according to monitored values of concentrations C_t: C_1, C_2, …, C_{24} (t = 1, 2, …, 24h).

Let $t_0 = 24h \rightarrow C_0 = C_{24}$

Thereby: $\alpha_1 = \frac{\ln \frac{C_1}{C_{24}}}{\ln \frac{24}{1}}; \alpha_2 = \frac{\ln \frac{C_2}{C_{24}}}{\ln \frac{24}{2}};$

……

$\alpha_{23} = \frac{\ln \frac{C_{23}}{C_{24}}}{\ln \frac{24}{23}}; \alpha_{24} = \frac{\ln \frac{C_{24}}{C_{24}}}{\ln \frac{24}{24}} = \frac{0}{0} (= \infty) -$ undefined (eliminated)

(Because the value of C_{24} has been chosen as the initial coordinate with t_0 = 24h).

Based on the above formulas, 23 values of $\alpha_1, \alpha_2, \ldots, \alpha_{23}$ will be determined to calculate the forecasted value of concentration C_t according to the input value of C_{24}.

2.3 Applying random function theory to establish interpolation/extrapolation model to repair PM_{10} data absence at Nguyen Van Cu monitoring station (Duong Ngoc Bach, 2012)

2.3.1 Scientific basis of the model

Considering the automatically monitored data at a fixed monitoring station is a random function of time t, i.e. x = x(t).

Monitoring data series are extracted for each day, with the values of $x(t_1)$, $x(t_2)$, ... separated by a time interval $\Delta t = \tau_1$ = 1h. Thereby, let $\tau = k\tau_1$, with k = 1, 2, …, (n-1).

Then, the time structure function calculated from the monitored data series is determined by the formula:

$$D(\tau) = D(k\tau_1) = \frac{1}{n-k} \sum_{i=1}^{n-k} (x_{i+k} - x_i)^2 \quad (11)$$

where xi and xi+k are monitored values at time ti and ti+k, respectively.

The interpolation/extrapolation formula is:

$$x(t_2) = x(t_1) \pm \sqrt{D_x(\tau)} \quad (12)$$

where the sign (+) happens when is increasing, while the sign (-) happens when is decreasing.

Apply the above formulas to interpolate or extrapolate using sliding method: to reduce the errors of interpolation or extrapolation, when interpolate or extrapolate the value x(t + τ) from x(t) with a relatively large τ, in here use the sliding method for interpolation or extrapolation.

For example, extrapolate $x(t_2)$ from $x(t_1)$ according to formula (12) with τ = 1h; then, take the extrapolated value $x(t_2)$ as input to extrapolate $x(t_3)$; …

The relative error of interpolation/extrapolation:

$$\text{MAPE} = \frac{1}{n} \sum_{1}^{n} \frac{|\text{actual value} - \text{interpolated/extrapolated value}|}{\text{actual value}} \quad (13)$$

2.4 Model of pollutants interacting with meteorological factors applied in Thailand (Pongpiachan and Paowa 2015)

2.4.1 Scientific basis of model

Materials and Method have been described thoroughly in the article (Pongpiachan and Paowa, 2015), only used formulas are taken into account.

Time Series Approach

Autocorrelation plot (Box and Jenkins) is a widely employed model for evaluating randomness in a data set and thus can be applied to investigate the randomness of OPD and IPD with time.

Autocorrelation plots can be conducted as follows. Firstly, vertical axis represents autocorrelation coefficient, which can be calculated by using the Eq.14.

$$R_h = \frac{C_h}{C_0} \quad (14)$$

where R_h is autocorrelation coefficient of patient number (i.e., Male-IPD, Female-IPD, Male-OPD, and FemaleOPD) and ranges between –1 and +1. Note that C_h is autocovariance function, which can be described in Eq.15.

$$C_h = \frac{1}{N} \sum_{t=1}^{N-h} (Y_t - \overline{Y})(Y_{t+h} - \overline{Y}) \quad (15)$$

where N, t, h, Yt, , Y_{t+h}, stand for total number of patients, time, time lag, number of patients at time t, average of patient numbers, and number of patients at time t + h, respectively. In addition, C_0 is the variance function, which can be written as follows:

$$C_0 = \frac{\sum_{t=1}^{N}(Y_t - \overline{Y})^2}{N} \quad (16)$$

Secondly, horizontal axis represents time lag h (h = 1, 2, 3, …). Thirdly, the confidence bands have fixed width that depends on sample size and can be calculated by using the following formula:

$$\pm \frac{Z_{1-\alpha/2}}{\sqrt{N}} \quad (17)$$

where N is the sample size, Z is the cumulative distribution function of the standard normal distribution and α is the significance level.

To investigate the influence of trace gaseous concentrations and meteorological variables on hospital admissions for respiratory diseases, Male-OPD and Female-OPD were modeled as:

$$\text{Male-OPD} = a + bT + cWS + dSin(WD) + eCos(WD) + fCO + gNOx + hSO2 + O3 \quad (18)$$

$$\text{Female-OPD} = a + bT + cWS + dSin(WD) + eCos(WD) + fCO + gNOx + hSO2 + O3 \quad (19)$$

Multiple linear regressions can establish the relative predictive importance of the independent parameters on the dependent variables. The analyses were performed using the SPSS 13.0 software for Microsoft Windows with the 'stepwise' MLRA method.

General Population Exposure of Outdoor Activities to PM_{10} and $PM_{2.5}$

To assess the health risks associated with general population exposure to both PM_{10} and $PM_{2.5}$ during the outdoor activities, an ILPE model was employed and defined as:

$$ILPE = C \times IR \times t \times EF \times ED \quad (20)$$

ILPE = Incremental lifetime particulate matter exposure (g)

C = PM_{10} and $PM_{2.5}$ concentrations (μg/m^3)

IR = Inhalation rate (m^3/h) t = Daily exposure time span (6 h/d, for two shifts)

EF = Exposure frequency (250 d/year [a] , upper-bound value)

ED = Exposure duration (25 years [a], upper-bound value)

Note: [a] Adapted from Human Health Evaluation Manual (US EPA, 1991).

3. Results and Discussion

3.1 Main results

The main results are calculated based on applying Pham Ngoc Ho's daily forecasting method for air pollutant index based on semi-empirical statistical theory.

Corrected forecasting equations of parameters' indexes

The coefficients a, b, c, d, e, g and h that are determined for each parameter with regression coefficient R^2 ranging from 0.83 - 0.99 from the consecutive monitored data in 2017 at Nguyen Van Cu station, Hanoi, Vietnam, and corrected forecast equations $q_c^*(t)$ with the corresponding regression coefficients as below:

For PM_{10} index:

$$q_c^*(t) = \frac{P}{P_0}\times\frac{f_0}{f}\times\frac{u_0}{u}\times\frac{T_0}{T}\times q_0(avg24h)\left[29.08\left(\frac{t}{t_0}\right)^6 - 105.27\left(\frac{t}{t_0}\right)^5 + 143.93\left(\frac{t}{t_0}\right)^4 - 92.314\left(\frac{t}{t_0}\right)^3 + 27.788\left(\frac{t}{t_0}\right)^2 - 3.3239\left(\frac{t}{t_0}\right) + 0.3523\right] + \alpha(t) \quad (21)$$

For NO_2 index:

$$q_c^*(t) = \frac{P}{P_0}\times\frac{f_0}{f}\times\frac{u_0}{u}\times\frac{T_0}{T}\times q_0(avg24h)\left[4.5055\left(\frac{t}{t_0}\right)^6 - 12.002\left(\frac{t}{t_0}\right)^5 + 13.066\left(\frac{t}{t_0}\right)^4 - 9.0078\left(\frac{t}{t_0}\right)^3 + 4.1843\left(\frac{t}{t_0}\right)^2 - 0.7571\left(\frac{t}{t_0}\right) + 0.1152\right] + \alpha(t) \quad (22)$$

For CO index:

$$q_c^*(t) = \frac{P}{P_0}\times\frac{f_0}{f}\times\frac{u_0}{u}\times\frac{T_0}{T}\times q_0(avg8h)\left[26.528\left(\frac{t}{t_0}\right)^6 - 93.889\left(\frac{t}{t_0}\right)^5 + 127.36\left(\frac{t}{t_0}\right)^4 - 82.514\left(\frac{t}{t_0}\right)^3 + 25.561\left(\frac{t}{t_0}\right)^2 - 3.143\left(\frac{t}{t_0}\right) + 0.1335\right] + \alpha(t) \quad (23)$$

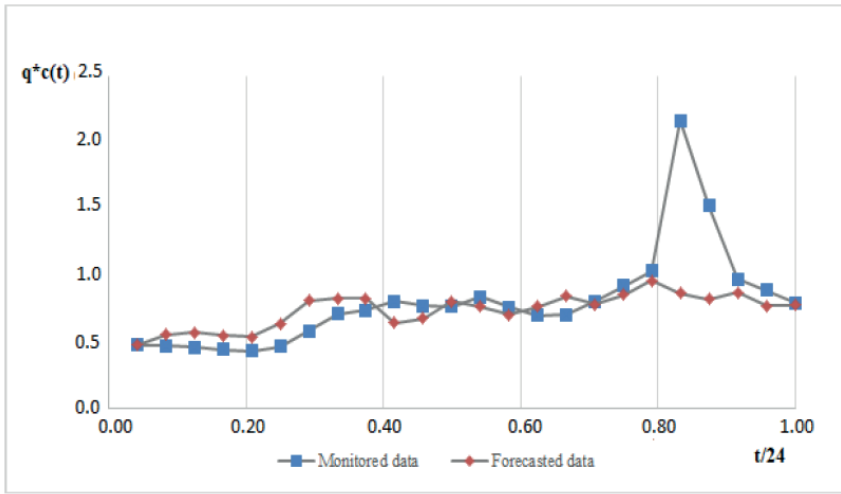
For O_3 index:

$$q_c^*(t) = \frac{P}{P_0}\times\frac{f_0}{f}\times\frac{u_0}{u}\times\frac{T_0}{T}\times q_0(avg8h)\left[-51.15\left(\frac{t}{t_0}\right)^6 + 166.88\left(\frac{t}{t_0}\right)^5 - 207.02\left(\frac{t}{t_0}\right)^4 + 121.16\left(\frac{t}{t_0}\right)^3 - 33.448\left(\frac{t}{t_0}\right)^2 + 3.6806\left(\frac{t}{t_0}\right) + 0.2175\right] + \alpha(t) \quad (24)$$

In which, α(t) is the correction factor, corresponding to the times of the day and calculated by the formula (4).

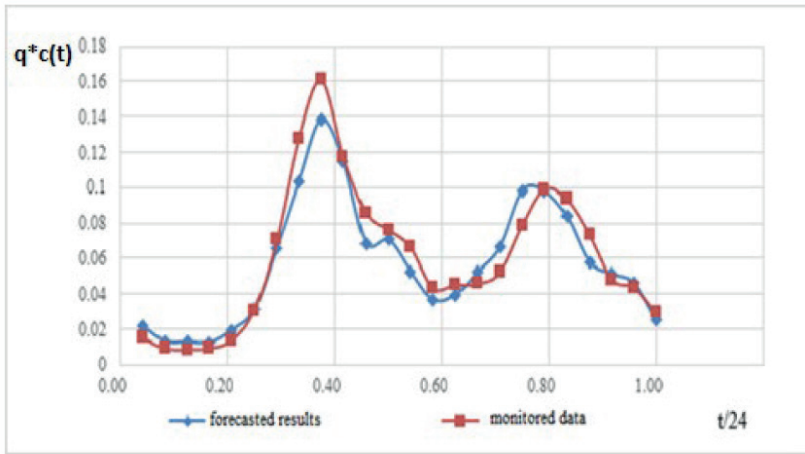
Corrected forecast results of parameters' indexes over time t

Some results of the forecasted values in comparison with the actual monitored values from Nguyen Van Cu monitoring station at standardized time points of t/24 in a specific day/month are illustrated in Figure 1 – 4. In which, monitored and corrected forecasting index of PM_{10} on December 2nd, 2018 from equation (21) is presented in Figure 1; monitored and corrected forecasting index of NO_2 on July 25th, 2018 from equation (22) is shown in Figure 2; that of CO on July 19th, 2018 from equation (23) and O3 on April 9th, 2018 from equation (24) are described in Figure 3 and Figure 4 respectively.



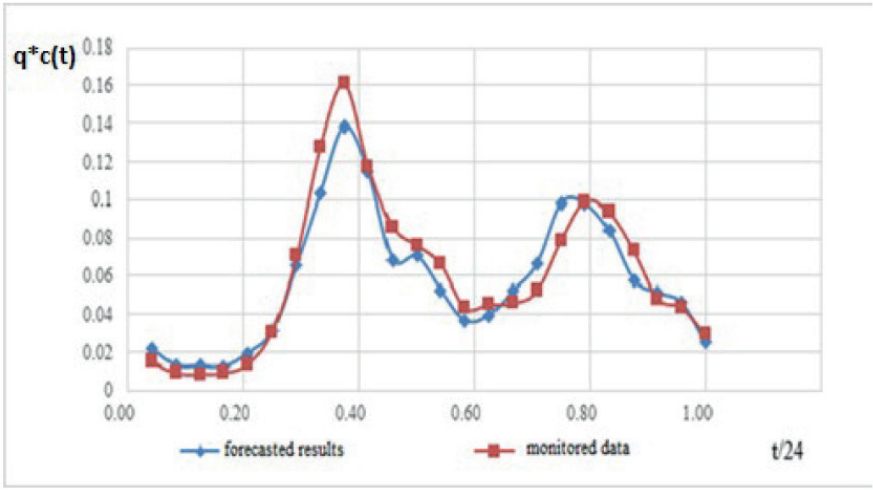
$$q_c^*(t) = \frac{C(t_i)}{C^*(1h)},\ C^*_{PM_{10}} = 210\ \mu g/m^3\ \text{(MONRE 2013)}$$

Figure 1. Monitored and corrected forecasting index of PM_{10} on 2nd December, 2018 from equation (21)



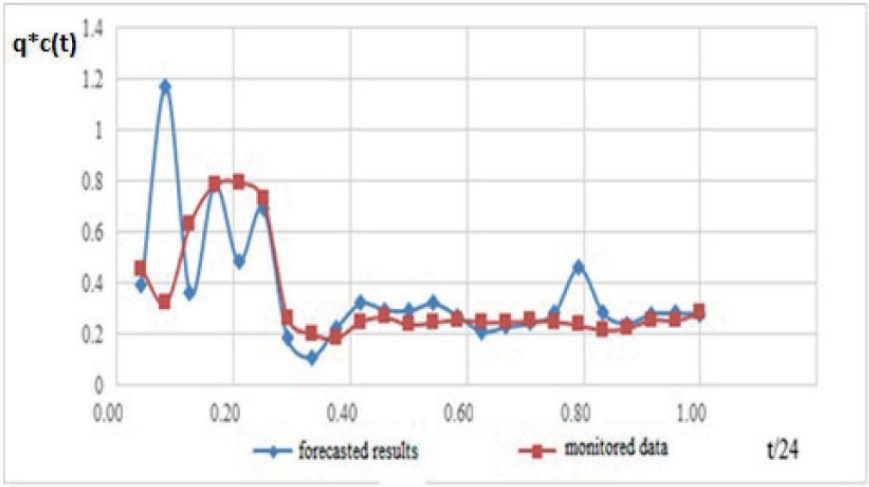
$$C^*_{CO} = 30{,}000\ \mu g/m^3\ \text{(MONRE 2013)}$$

Figure 2. Monitored and corrected forecasting index of NO_2 on 25th July, 2018 from equation (22)



$C^*_{CO} = 30{,}000\ \mu g/m^3$ (MONRE 2013)

Figure 3. Monitored and corrected forecasting index of CO on 19th July, 2018 from equation (23)



$C^*_{O_3} = 200\ \mu g/m^3$ (MONRE 2013)

Figure 4. Monitored and corrected forecasting index of O3 on 9th April, 2018 from equation (24)

3.2 Other results

Results drawn from 3 other models are used to compare with the main results.

3.2.1. Results calculated from Nguyen Van Cu monitoring station's data (2017) applied Hanna SR Statistical Model

Result of α_t of pollutants on 6th December 2017 is presented in Table 1, results of forecasted values with relative errors is presented in Table 2, and graphs of monitored data and forecasted values are shown in Figure 5 – 8.

Table 1. The value of αt of pollutants at time t on 6th December 2017, calculated based on Hanna SR's model

	PM_{10}	NO_2	CO	O_3
α_1	0.016	0.465	0.043	-0.052
α_2	-0.009	0.589	0.021	-0.039
α_3	0.167	-0.401	0.081	-0.034
α_4	0.068	0.763	0.081	0.016
α_5	0.056	0.837	0.083	-0.039
α_6	0.053	1.108	0.035	-0.247
α_7	-0.036	1.395	0.086	-1.313
α_8	0.250	-0.347	0.478	-1.549
α_9	0.531	0.126	0.853	-1.683
α_{10}	-0.115	-0.295	-0.035	-2.062
α_{11}	-0.190	-0.139	0.152	-1.140
α_{12}	-0.314	-0.092	0.193	-1.377
α_{13}	-0.285	0.169	0.307	-2.652
α_{14}	-0.405	0.415	0.495	-3.575
α_{15}	-0.424	0.654	0.936	-1.929
α_{16}	-0.122	0.781	1.094	-0.867
α_{17}	-0.032	0.699	1.394	-1.595
α_{18}	0.261	0.585	1.589	-1.950
α_{19}	1.602	1.361	2.714	-2.897
α_{20}	2.119	1.560	3.288	-3.870
α_{21}	2.521	1.213	3.355	-5.247
α_{22}	0.996	1.024	4.370	-4.246
α_{23}	0.856	2.059	5.979	-4.810

Table 2. Forecasted values and relative error of each pollutant at time t in 2018, calculated based on Hanna SR's model

Time t (hour)	PM_{10} (µg/m³)		NO_2 (µg/m³)		CO (µg/m³)		O_3 (µg/m³)	
	Forecasted value	Error	Forecasted value	Error	Forecasted value	Error	Forecasted value	Error
1	70,52	0,42	1342,18	2,65	2462,40	0,05	13,91	0,65
2	65,43	0,54	1322,26	2,66	2261,46	0,03	14,88	0,34
3	94,80	0,20	132,85	0,59	2543,36	0,44	15,27	0,40
4	75,68	0,76	1200,72	2,85	2483,73	0,56	16,87	0,36
5	73,08	0,58	1135,98	2,26	2448,46	0,74	15,42	0,42
6	72,09	0,47	1421,29	2,40	2255,84	0,15	11,63	0,22
7	64,05	0,62	1704,80	2,63	2387,90	0,20	3,25	0,72
8	88,11	0,57	208,92	2,95	3630,26	0,02	2,99	0,80
9	112,69	0,42	346,11	5,42	4958,64	0,58	3,14	0,78
10	60,57	0,75	236,09	3,37	2083,42	0,24	2,69	0,81
11	57,76	0,57	274,22	3,49	2419,67	0,08	6,73	0,70
12	53,88	0,82	286,86	3,23	2456,03	0,84	6,31	0,85
13	56,22	0,77	339,14	4,02	2592,92	2,11	3,22	0,95
14	53,83	0,79	382,39	5,10	2805,84	1,37	2,39	0,97
15	54,87	0,82	415,69	5,78	3334,68	0,89	6,62	0,94
16	63,73	0,81	419,59	4,41	3347,56	0,71	11,53	0,90
17	66,23	0,81	389,03	0,12	3474,60	0,32	9,45	0,93
18	72,20	0,83	361,72	0,02	3393,44	0,56	9,35	0,92
19	97,38	0,78	420,15	6,43	4049,63	0,42	8,33	0,88
20	98,56	0,60	406,31	6,63	3912,08	0,17	8,09	0,84
21	93,78	0,58	359,50	5,55	3362,45	0,04	8,13	0,73
22	73,04	0,13	334,23	5,58	3142,33	0,03	11,32	0,66
23	69,46	0,08	333,74	0,16	2770,81	0,29	13,35	0,49

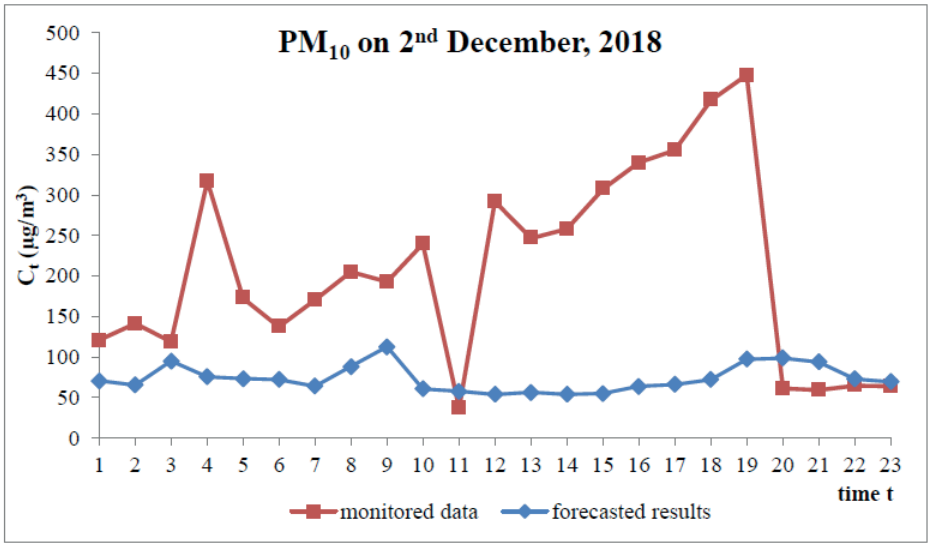


Figure 5. Monitored and forecasted results of PM_{10} on 2nd December, 2018 from equation (9), based on Hanna SR's model

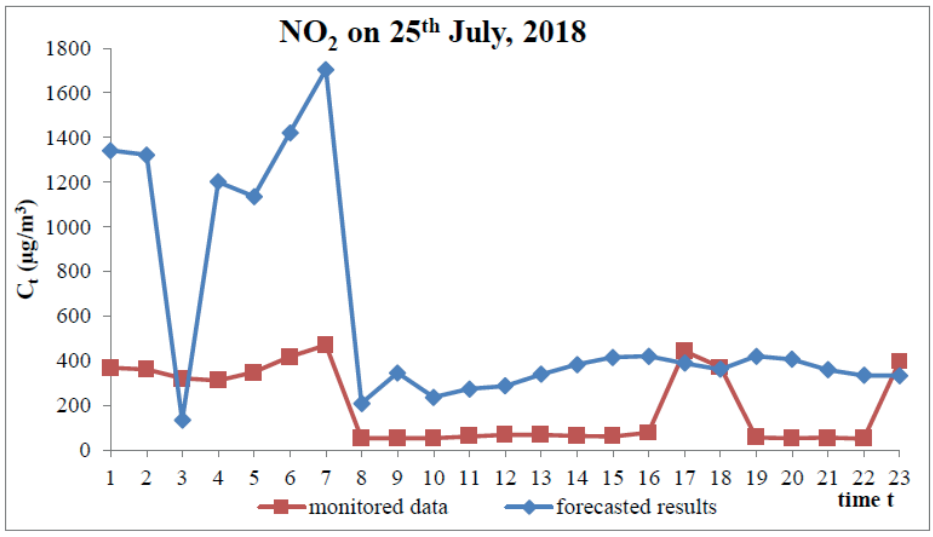


Figure 6. Monitored and forecasted results of NO_2 on 25th July, 2018 from equation (9), based on Hanna SR's model

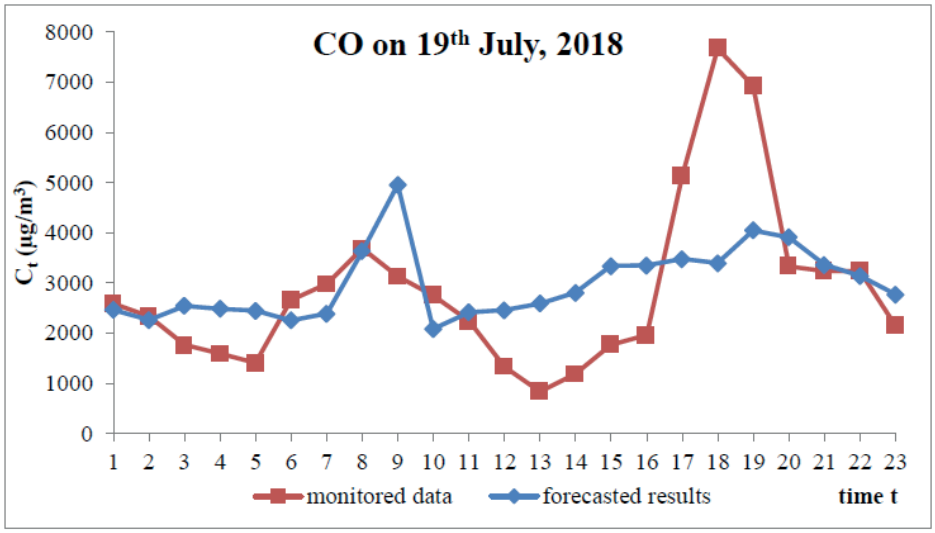


Figure 7. Monitored and forecasted results of CO on 19th July, 2018 from equation (9) based on Hanna SR's model

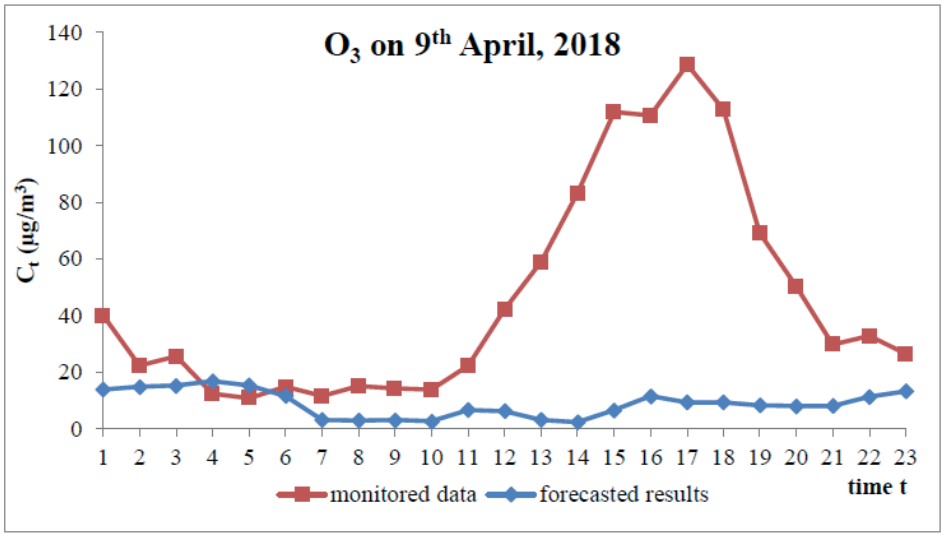


Figure 8. Monitored and forecasted results of O_3 on 9th April, 2018 from equation (9), based on Hanna SR's model

3.2.2. Results calculated for Nguyen Van Cu monitoring station applied random function theory to establish interpolation/ extrapolation model to repair PM_{10} data absence (Duong Ngoc Bach, 2012)

With the purpose of comparing results with ones drawn from other models, here are a number of relevant main results, presented in Table 3 and graphs illustrating the structure function (Figure 9), the extrapolation (forecast) results in 24 hour-a-day (Figure 10).

Table 3. Average error of data extrapolation model (Duong Ngoc Bach, 2012)

Month (Data extrapolation)	Average error	Month (Data extrapolation)	Average error
January 2012	13%	July 2012	27%
February 2012	13%	August 2012	26%
March 2012	14%	September 2012	21%
April 2012	18%	October 2012	14%
May 2012	28%	November 2012	18%
June 2012	22%	December 2012	-

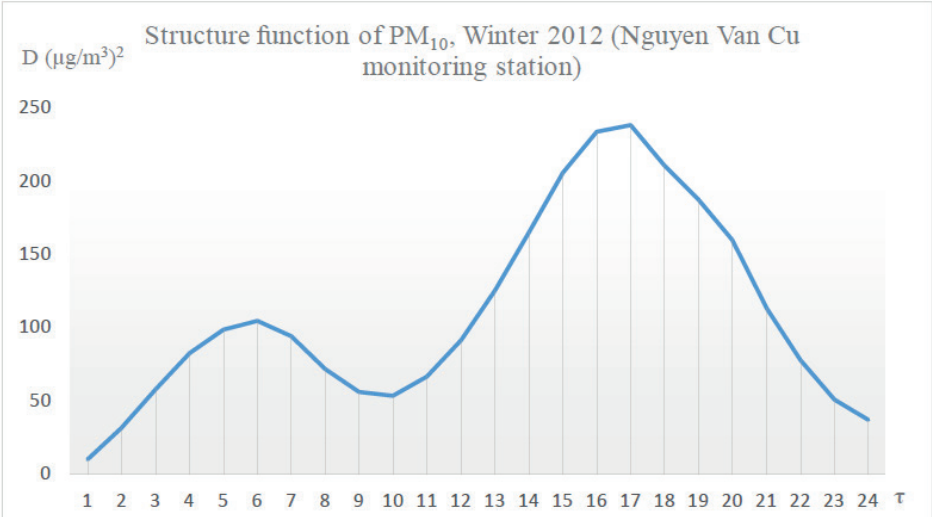
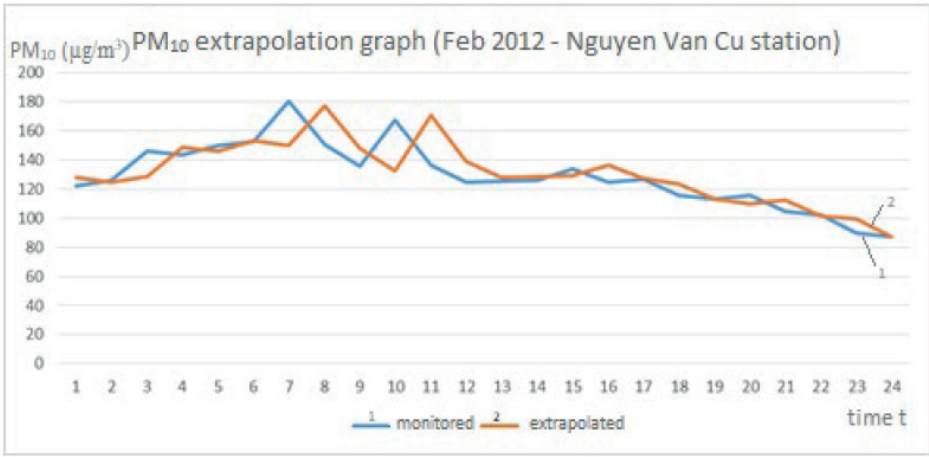


Figure 9. Structure function of PM_{10}, Winter 2012 (Nguyen Van Cu monitoring station) (Duong Ngoc Bach, 2012) from equation (11)

Figure 10. The extrapolation results of PM_{10} (February 2012) - Nguyen Van Cu station (Duong Ngoc Bach, 2012) from equation (12)



3.2.3. Results applied Model of pollutants interacting with meteorological factors applied in Thailand (Pongpiachan and Paowa, 2015)

Trace gaseous species, meteorological parameters, numbers of Male-OPD, Female-OPD, Male-IPD and FemaleIPD were identified successfully throughout the sampling campaign (n = 2,312). Table 4 summarizes the average concentrations of all air quality parameters measured at TMCS coupled with hospital admissions monitored at NHCM. To assess the impact of haze episodes on meteorological parameters and patient numbers, the data were separated into two groups: *"haze period"* and *"nonhaze period"*. Because March is the month with the most severe haze episodes in Chiang-Mai, data collected in March of each year from 2007 to 2013 were assigned to the *"haze period"* and data collected from January 1, 2007 to April 30, 2013, excluding March data, to the *"non-haze period"*. The atmospheric concentrations of $PM_{10\text{-"haze period"}}$ varied from N.D. to 128 µg/m^3 with an average of 72.5 ± 26.2 µg/m^3 , and the concentrations of $PM_{10\text{-"non-haze period"}}$ ranged from N.D. to 118 µg/m3 with an average of 34.7 ± 19.5 µg/m^3 . These two groups differ significantly in average PM_{10} concentrations based on the two-sample t-Test ($p < 0.05$). Statistical descriptions of OPD and IPD patient numbers at NHCM during 2007–2013 are displayed by gender and age in Table 5. The ANOVA results indicate some significant differences among age groups in OPD and IPD patient numbers. For instance, Male-OPD and Female-OPD show the highest values of 21.3 ± 10.9 ($p < 0.005$) and 27.8 ± 14.0 ($p < 0.005$) at age group 0–14 and age group 15–59, respectively.

For the age group 0–14, Male-IPD and Female-IPD displayed the maximum values of 2.1 ± 1.8 and 1.5 ± 1.4, respectively. The atmospheric concentrations of CO, NO_x, SO_2 and O_3 were assessed for the two periods. All trace gas concentrations detected during the haze episode were significantly higher than those of the *non-haze period*, as displayed in Table 4 ($p < 0.05$).

3.3 Discussion

PM_{10} index:

Figure 1 shows that the efficiency in each time t from 1h to 18h and from 22h to 24h is 95-97%, in the duration 20h-21h is 40.1-54%. Daily relative error average (24-hour average) is 0.079 with the efficiency of the model is 92.1%.

NO_2 index:

Figure 2 shows that error at duration t=2h-10h; 13h-24h is within a range 0.05-0.21 with the efficiency of 79-95%. At the time t of 1h, 11h, 12h and 14h, the error is around 0.4-0.6 with the efficiency of 40-60%. The daily error average of the index is 0.019, equivalent to the model's efficiency of 98.1%.

CO index:

Figure 3 shows that error of every time is around 0.01-0.16, meaning that the model's efficiency is 84-99%. 8-hour error average is 0.106, meaning that the model's efficiency is 89.4%.

O_3 index:

Figure 4 shows that error in each time t is within a range of 0.002-0.22, meaning that the model's efficiency is 78-99.8%, but there is one exception at t=2h with the error of 0.61. Daily 8-hour error average is 0.0046, equivalent to the efficiency of 99.54%.

In each case, the error between $q_c^*(t)$ at the time t and $q_m(t)$ is: $\frac{2}{24}$ = 8.33% for Figure 1 with 2 values out of 24 forecasted values; $\frac{4}{24}$ =16.67% for Figure 2; $\frac{1}{24}$ = 4.17% for Figure 3 and $\frac{3}{24}$ = 12.5% for Figure 4. These errors are minor compared to the total of 24 forecasted values.

The relatively large errors occurring at some standardized time points of t/24 in Figure 1, Figure 2 and Figure 4 are representations of the results of the comparison between actual monitored values from Nguyen Van Cu monitoring station and the forecasted values in a specific day/month. These errors can be explained as follows: the actual data used has been taken into account with the combined effect of many factors such as the transition between pollutants according to chemical reactions, increasing traffic volume in the mornings and afternoons, the influence of meteorological factors on the atmospheric state (stable, unstable and equilibrium), etc. So, if the data provided to us are of high accuracy, then the forecast error will also depend on the error of meteorological forecast which is given by the Vietnam National Center for Forecasting on the media. Therefore, the forecast results of the model are inevitably affected.

By using the corrected forecasting equations (21), (22), (23) and (24) for 7 consecutive days of each month from January to December in 2018, the results show that the efficiency of our corrected forecasting model over time t for a typical day in each month is 75-95%. Our model also works effectively in case of 8-hour average and 24-hour average with the efficiency of 85-98%. Based on these strong evidences, we have concluded that our corrected forecasting model has high efficiency and could be applied successfully in reality.

The forecasted results based on Hanna S.R's model are illustrated by the graphs Figure 5 - 9. The relative error at t = 1 - 23h corresponding to the pollutants is 0.02 – 6.63 (NO_2), 0.02 – 2.11 (CO), 0.22 – 0.95 (O_3) and 0.08 – 0.83 (PM_{10}).

The forecast results of PM_{10} dust at Nguyen Van Cu monitoring station based on the interpolation/extrapolation model are presented in Figure 10 - 11, the relative error at t = 1 - 24h has corresponding values from 0.13 to 0.28.

Results of the model of chemicals and dust interacting with meteorological elements (Thailand) have high accuracy, which are shown in Table 4 and Table 5. The ANOVA results revealed a significant increase in hospital walk-ins and admissions for both genders in the < 15 years group ($p < 0.005$). MLRA revealed the significantly highest impacts of CO on hospital walk-ins for both genders. The predicted ILPE of PM10 showed the highest values for both genders during the *"haze-episode"* in 2007, with average values of 3.338 ± 0.576 g and 1.838 ± 0.317 g for male and female outdoor workers, respectively, over exposure duration of 25 years (Pongpiachan and Paowa, 2015).

4. Conclusion

The authors applied semi-empirical statistical model, random function theory, and model of chemicals and dust interacting with meteorological factors to predict pollution parameters in the air layer close to the ground.

The above models have different approaches and calculation formulas for chemical parameters (NO_2, SO_2, CO, O_3, dust and meteorology) that have been cited for comparison such as Pongpiachan and Paowa's, applied in Thailand; Duong Ngoc Bach's, Pham Ngoc Ho's, Hanna SR's, applied in Vietnam. The results obtained from these models are different, but they all achieve high accuracy for practical application, except for Hanna SR's model which has delivered a relatively great error. However, the model we use has an outstanding advantage of forecasting the air pollution index according to the daily forecast of meteorological factors on the mass media. This is a new approach that has never been reportedly applied to any contemporary modelling.

Table 4. Statistical description of trace gaseous species, meteorological parameters, OPD and IPD patient numbers at Nakornping Hospital, Chiang-Mai province, 2007– 2013 (Pongpiachan and Paowa, 2015)

	CO	NO_x	SO_2	O_3	PM_{10}	RH	P	WS	WD	T	Male-OPD	Female-OPD	Male-IPD	Female-IPD
Non-Haze Period (2007-2013)	0.450 ± 0.210	11.2 ± 5.97	2.86 ± 3.30	15.3 ± 10.4	34.7 ± 19.5	66.2 ± 19.5	3.27 ± 8.70	11.7 ± 4.68	208 ± 124	26.4 ± 2.82	47.8 ± 22.4	53.0 ± 25.1	5.18 ± 2.68	3.90 ± 2.29
Haze Period (2007-2013)	0.704 ± 0.272	16.3 ± 5.86	4.66 ± 5.97	18.6 ± 12.6	72.5 ± 26.2	55.9 ± 14.4	0.535 ± 2.57	11.4 ± 4.76	193 ± 71.8	27.1 ± 2.08	47.8 ± 20.2	54.1 ± 23.7	5.50 ± 2.53	4.41 ± 2.27
t-Test ($p < 0.05$)	S	S	S	S	S	S	S	NS	S	S	NS	NS	NS	S

CO: Carbon monoxide [ppm], NO_x: Nitrogen oxides [ppb], SO_2: Sulfur dioxide [ppb], O_3: Ozone [ppb], PM_{10}: Particulate Matter of 10 Microns in diameter or smaller, RH: Relative humidity [%], P: Daily Precipitation [cm], WS: Wind speed [m/s], WD: Wind direction, T: Temperature [°C], Male-OPD: Number of male walk-in patients, Female-OPD: Number of female walk-in patients, Male-IPD: Number of admitted male patients, Female-IPD: Number of admitted female patients, S: Significant, NS: Not Significant.

Table 5. Statistical description of OPD and IPD patient numbers by gender and age at NHCP from 2007 to 2013 (Pongpiachan and Paowa, 2015)

Age Range (Years)	0-14		15-59		60-74		75+		F-Value (p<0.005)	Statistical Significance
	Aver	Stdev	Aver	Stdev	Aver	Stdev	Aver	Stdev		
Male-OPD	21.3	10.9	18.7	9.5	6.1	4.2	3.5	2.7	3137	S
Female-OPD	17.3	8.8	27.8	14.0	5.8	4.2	3.7	2.9	3856	S
t-Value (p<0.005)	13.7		-25.9		2.4		-2.4			
Statistical Significance	S		S		NS		NS			
Male-IPD	2.1	1.8	1.3	1.2	1.0	1.0	0.9	1.0	410	S
Female-IPD	1.5	1.4	0.80	0.92	0.78	0.92	0.86	1.0	237	S
t-Value (p<0.005)	12.6		15.9		7.8		1.4			
Statistical Significance	S		S		S		NS			

Aver: Average, Stdev: Standard Deviation.

Acknowledgement

The authors would like to thank the Environmental Monitoring Center of Vietnamese Environment Administration for providing us the monitoring data of 2017 and 2018 for this research.

References

Berliand ME. Forecast and Atmospheric Contamination. Hydro meteorological Publishing House. Leningrad, Russia. 1985 (in Russian).

Duong Ngoc Bach. Applying random function to establish interpolation, extrapolation models to repair the absence of dust PM10 data series at automatic air quality monitoring stations in Hanoi. VNU University of Science, project code: TN-10-56. Hanoi, Vietnam. 2012 (in Vietnamese).

Duong Ngoc Bach. Simulation of variation and transporting process of PM_{10} dust in the air in Hanoi. Doctoral Thesis in Environmental Science. Hanoi, Vietnam. 2016 (in Vietnamese).

GRIMM Aerosol Technik. https://www.grimm-aerosol.com/products-en/environmental-dust-monitoring/approved-pm-monitor/edm180/. Accessed November 15th, 2019.

Hanna SR. Review of Atmospheric Diffusion Models for Regulatory Application. WHO Technical Note 177, 1982; 1-37.

HORIBA Process and Environment. Air Pollution Monitor AP-370 Series. https://static.horiba.com/fileadmin/Horiba/Products/Process_and_Environmental/Ambient/Brochures/AP-370_bro_E_HRE-2858G.pdf. Accessed November 15th, 2019.

Kazakevits DI. Basics of random function theory applying in meteorology and hydrology. Leningrad, Russia. 1971 (in Russian).

Ministry of Natural Resources and Environment of Vietnam (MONRE). Vietnam Technical Regulation on Ambient Air Quality QCVN 05:2013/MONRE. Hanoi, Vietnam. 2013 (in Vietnamese).

Pasquill F. Atmospheric Diffusion: The Dispersion of Windborne Material from Industrial and Other Sources. 2nd ed., John Wiley & Sons. New York, USA. 1974.

Pham Ngoc Ho. Applying mathematics in environmental science. Vietnam National University Press 2016, code 186-KHTN-2016. Hanoi, Vietnam. 2016 (in Vietnamese).

Pham Ngoc Ho. Forecasting air pollutant index model based on semi-empirical statistical theory. Proceedings of the International Conference on Environment and Sustainable Development in Mineral Resource Extraction. Vietnam Academy of Science and Technology Press 2017; 3-12; ISBN: 978-604-913-623-8.

Pham Ngoc Ho, Trinh Thi Thanh, Dong Kim Loan, Pham Thi Viet Anh. Basics of Air and Water Environment. Vietnam National University Press 2011; 19 (in Vietnamese).

Pongpiachan S. FTIR spectra of organic functional group compositions in PM2.5 collected at Chiang-Mai City, Thailand during the haze episode in March 2012. J. Appl. Sci. 2014; 14(22): 2967-2977.

Pongpiachan S, Hattayanone M, Suttinun O, Khumsup C, Kittikoon I, Hirunyatrakul P, Cao J. Assessing human exposure to PM10-bound polycyclic aromatic hydrocarbons during fireworks displays. Atmospheric pollution research 2017; 8(5): 816-827.

Pongpiachan S, Kositanont C, Palakun J, Liu S, Ho KF, Cao J. Effects of day-of-week trends and vehicle types on PM2.5-bounded carbonaceous compositions. Science of the Total Environment 2015; 532; 484-494.

Pongpiachan S, Paowa T. Hospital out-and-in-patients as functions of trace gaseous species and other meteorological parameters in Chiang-Mai, Thailand. Aerosol and Air Quality Research 2015; 15(2): 479-493.

Schonoor JL. Environmental Modeling: Fate and Transport of Pollution in Water, Air and Solid. J. Wiley, New York, USA. 1996.

Tran Thi Thu Huong. Research of daily variation, interpolation and extrapolation of CO and PM10 in several permanent automatically monitoring stations in Vietnam. Doctoral Thesis in Environmental Science. Hanoi, Vietnam 2017 (in Vietnamese).