Atmospheric Environment 71 (2013) 36–43

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia

Jasim M. Rajab [a], M.Z. MatJafri [b], H.S. Lim [b,*]

[a] Mosul University, Physics Department, College of Science, Mosul, Iraq
[b] School of Physics, Universiti Sains Malaysia, 11800 Penang, Malaysia

HIGHLIGHTS

► A model for prediction of the value of atmospheric columnar ozone (O_3).
► The O_3 was negatively correlated with CH_4, H_2O vapour, RH, and MSP.
► The O_3 was positively correlated with CO, AST, SSKT, and AT.
► Close agreement between the predicted and the AIRS observed data for columnar O_3.
► AIRS data are suitable used to investigate the impact of atmosphere parameters.

ARTICLE INFO

Article history:
Received 9 October 2012
Received in revised form
8 January 2013
Accepted 10 January 2013

Keywords:
Ozone (O_3)
Regression analysis
Principal component analysis
Atmosphere infrared sounder (AIRS)

ABSTRACT

This study encompasses columnar ozone modelling in the peninsular Malaysia. Data of eight atmospheric parameters [air surface temperature (AST), carbon monoxide (CO), methane (CH_4), water vapour (H_2O_{vapour}), skin surface temperature (SSKT), atmosphere temperature (AT), relative humidity (RH), and mean surface pressure (MSP)] data set, retrieved from NASA's Atmospheric Infrared Sounder (AIRS), for the entire period (2003–2008) was employed to develop models to predict the value of columnar ozone (O_3) in study area. The combined method, which is based on using both multiple regressions combined with principal component analysis (PCA) modelling, was used to predict columnar ozone. This combined approach was utilized to improve the prediction accuracy of columnar ozone. Separate analysis was carried out for north east monsoon (NEM) and south west monsoon (SWM) seasons. The O_3 was negatively correlated with CH_4, H_2O_{vapour}, RH, and MSP, whereas it was positively correlated with CO, AST, SSKT, and AT during both the NEM and SWM season periods. Multiple regression analysis was used to fit the columnar ozone data using the atmospheric parameter's variables as predictors. A variable selection method based on high loading of varimax rotated principal components was used to acquire subsets of the predictor variables to be comprised in the linear regression model of the atmospheric parameter's variables. It was found that the increase in columnar O_3 value is associated with an increase in the values of AST, SSKT, AT, and CO and with a drop in the levels of CH_4, H_2O_{vapour}, RH, and MSP. The result of fitting the best models for the columnar O_3 value using eight of the independent variables gave about the same values of the R ($\approx$0.93) and R^2 ($\approx$0.86) for both the NEM and SWM seasons. The common variables that appeared in both regression equations were SSKT, CH_4 and RH, and the principal precursor of the columnar O_3 value in both the NEM and SWM seasons was SSKT.

1. Introduction

Ozone (O_3) is unique among pollutants because it is not emitted directly into the air, and its results from complex chemical reactions in the atmosphere. However, it has dramatically different effects depending on where O_3 resides. It can harm or protect life on earth (Struijs et al., 2010; Mintzer and Miller, 1987). This is the main reason why ozone is such a serious environmental problem that is difficult to predict and control. The O_3 has been identified as one of the primary pollutants that degrade air quality, and it is naturally present in our atmosphere. Because of its ability to absorb infrared radiation it considered an essential and important greenhouse gas, and its presence both at the ground-level and in the earth's upper atmosphere (Freijer et al., 2002). Stratospheric ozone is considered

* Corresponding author.
E-mail address: hslim@usm.my (H.S. Lim).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.01.019

to be beneficial for humans and other life forms because it creates a protective shield by absorbing some of the sun's biologically harmful ultraviolet (UV)-B radiation. In humans, increased exposure to UV radiation can lead to an increase in the occurrence of skin cancer, cataracts and immune-system impairment (Bian et al., 2007). Due to its involvement in a number chemical and biological process in the surface as well as at the troposphere, the UV radiation has been critically important for humans and other living beings (Hassanzadeh et al., 2008).

At the ground-level, O_3 is a harmful pollutant that causes damage to lung tissue, plants and other living systems, and it not emitted directly into the air but is formed by the chemical reaction between the volatile organic compounds (VOCs) and the nitrogen oxides (NOX) (combination of NO and NO_2) in the presence of sunlight and heat. The VOCs are emitted from various sources, including motor vehicles and other industrial sources. Ground-level ozone forms easily in the atmosphere, particularly in warm sunny urban areas (Morris et al., 2006). Weather patterns play a major role in terminating episodes of high ozone concentrations. Wind speed also plays a major role in the formation of ozone. The ozone precursor's species are directly determine by the meteorological processes whether it contained locally or transported downwind with the resulting ozone. Ozone formation is most conducive during cloudless, dry warm days with low wind speeds; these conditions most often occur during high-pressure systems (Duenas et al., 2002; NAP, 1991; Bossioli et al., 2007) Depending on the wind direction, ozone and the precursor pollutants that form ozone can also be transported hundreds of kilometres away (DOE, 2004). Parameters which may impact ozone values have been investigated in several studies (Rajab et al., 2011; Abdul-Wahab, 2001, 2005; Al-Alawi et al., 2008; Blankinship, 1996; Hassanzadeh et al., 2008; Sebald et al., 2000; Zhang et al., 2012).

Multiple regression analysis (MRA) is one of the most widely employed in the atmospheric sciences for expressing the dependence of a response variable on several independent (predictor) variables by fitting a linear equation to observed data. Regardless of its obvious success in many applications, however, when the independent variables are correlated with each other, the regression approach can face serious difficulties (Rajab et al., 2012; McAdams et al., 2000). The principal component analysis (PCA) is useful for mitigating the problem of multicollinearity. In addition, exploration of the relationships among the independent variables can be used to filter the data so that only the significant independent variables responsible for the dependent observation can be determined. PCA has been used in air quality studies to separate interrelationships into statistically independent basic components (Maenhaut et al., 1989; Vaidya et al., 2000).

The abundances of atmospheric gases over the past three decades have been measured using balloons, airplanes and sparsely distributed measurement sites. The measurements were mostly confined to the surface of the site, and unable to make continuous recordings of global variation over the long-term (Tiwari et al., 2005). Satellite remote sensing can provide continuous data with high spatial and temporal resolution (Dousset and Gourmelon, 2003; Illingworth et al., 2011). In addition, its good global coverage increased our capability to assess the influence of human activities on climate change and on the chemical composition of the atmosphere (Clerbaux et al., 2003).

The AIRS is one of the several instruments onboard the Earth Observing System (EOS), onboard NASA's Aqua Satellite, launched on May 4, 2002. AIRS is a continuously operating cross-track scanning sounder consisting of a telescope that feeds a scale spectrometer. The AIRS instrument views the atmospheric infrared spectrum in 2378 channels with a nominal spectral resolving power $\lambda/\Delta\lambda$ ranging from 1086 to 1570 covering more than 95% of the earth surface and returning about three million spectra daily in the 3.74–4.61 μm, 6.20–8.22 μm and 8.8–15.4 μm infrared wavebands at a nominal spectral resolution (Strow et al., 2003; Wang et al., 2012).

In this study, we combine the multiple regression method and PCA to obtain regression equations for total column of ozone with other measured ambient atmosphere parameters as predictor variables. The idea is to use results from the analysis of the retrieved gases in the atmosphere for the period (2003–2008) obtained from the AIRS data to develop regression equation for calculating the columnar ozone during NEM and SWM seasons over peninsular Malaysia. The particular aim is to relate the accurate columnar O_3 value to atmosphere parameters using satellite data. In addition, to investigate and analyses the influences of the atmosphere parameters in the columnar O_3 values using statistical methods.

2. Materials and methods

2.1. Study area and data collection

The study area of peninsular Malaysia is located from 99° to 105° longitude east and between 1° and 7° latitude north. The study area is north of Singapore, south of Thailand, and east of the Indonesian island of Sumatra. It comprises an area of 131,587 km^2 consisting of floodplain, coastal zones and highland (Fig. 1). The Titiwangsa Mountain form the backbone of the peninsula Malaysia, range from the Malaysia-Thai border in the north running approximately south-southeast over a distance of 480 km, and separating the western part from the eastern part (Suhaila and Jemain, 2007).

In peninsular Malaysia, as is characteristic for a humid tropical climate, the weather is warm and humid throughout the year with temperatures ranging from 20° to 32 °C, and recorded air temperatures of 38 °C are very rare. There are annual fluctuations of the mean temperature of roughly 1.5–2 °C. The lowest average monthly temperatures occur from November–January, and the highest average temperatures occur from April–May and July–August in most places (Dasimah, 2009). There is a monthly definite variation that coincides with the monsoons. The climate of peninsular Malaysia meaningfully influences by the monsoons and experiences two rainy seasons throughout the year, associated with the South west Monsoon (SWM) from May to August and the North east Monsoon (NEM) from November to February (Wong et al., 2009).

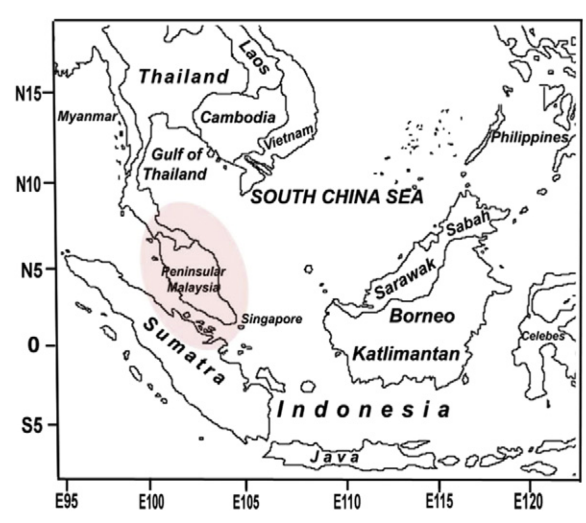

Fig. 1. The geographical feature of the study area.

The lightning and variable wind thunderstorms develop, during the inter-monsoon months (usually between April and October), causing substantial rainfall in each of the two transition periods, especially in the west coast states. The main factors that affect the rainfall distributions are the monsoon changes and the effects of topography (Suhaila and Jemain, 2009). Besides the contribution of the many regional pollutant sources, these monsoons have different influences on the atmospheric parameters, in terms of the effects on climate or the amounts of pollutants; they bring to Malaysia.

The main objective of the study is to develop regression models for column of ozone for NEM and SWM seasons over peninsular Malaysia by processing the satellite data for the retrieved atmosphere parameter. Seven years of satellite data were collected from January 2003 to December 2009. Six-year data (January 2003 to December 2008) were collected during this duration for analysis and development of the predictive column of O_3 regression models. The validations, comparisons and mapping of the data were conducted by using the last year, 2009. The satellite data were collected online from National Aeronautics and Space Administration (NASA), and the *in situ* data were collected from the Malaysian Meteorology Department (MMD). The various kinds of software that were used during the process include Statistical Package for Social Sciences (SPSS), Microsoft Excel (MS-Excel), and SigmaPlot 11.

There are four distinct processing phases for processing the AIRS PGEs (Product Generation Executive): Level-1A, Level-1B, Level-2 (L2) and Level-3 (L3). The single Level-2 PGE receive granules from Level-1B data and produce geolocated, calibrated cloud-cleared radiances and 2-dimensional and 3-dimensional retrieved physical quantities products (e.g., moisture, surface properties and temperature, ozone, carbon monoxide and methane profiles throughout the atmosphere). These geophysical estimates are designated Level 2 (Ye et al., 2007), which produces 240 granules of each of cloud-cleared radiances, standard and support products every day. Each product granule contains 6 min of data fields from 1350 retrievals laid out in an array of dimension 30×45, matching to the 30 AMSU footprints (cross-track) in each of 45 scansets (along-track) (Fishbein et al., 2007). The L3 data are created from the L2 data product by binning them in $1^\circ \times 1^\circ$ grids. Level-3 products are statistical summaries of geophysical parameters that have been temporally aggregated and spatially re-sampled from lower level data products (e.g., Level-2 data) (Pagano et al., 2006). The AIRS radiance data in the 9.6-μm band is used to retrieve column of ozone and ozone profiles for both day and night (including the polar night). The initial guess ozone profiles are obtained and given at 100 levels. Next, they are used in the physical retrieval algorithm, which discloses the geophysical parameters that provide the best counterpart for cloud-cleared regression in a subset of the AIRS channels (Pittman et al., 2009). The AIRS measures the total column of O_3 by (41) channels with an uncertainty estimate of 5% at the tropics and 5–40% at the poles. Because AIRS is an IR (thermal) instrument, it is more sensitive to the ozone distribution in the coldest portion of the atmosphere, in the proximity of the tropopause (Bian et al., 2007).

The Level-3 standard data are stored in the HDF-EOS4 format, and can be easily arranged in Microsoft excel. Then Statistical Package for Social Sciences (SPSS) software is used for analysis the data to find correlations between atmospheric parameters and predicted equations of columnar O_3 using MRA and PCA. The SigmaPlot 11 was used to correlate and validate the predicted and the observed columnar O_3, and to plot their values for comparison in 2009.

Generally, 84 monthly L3 ascending AIRX3STM $1^\circ \times 1^\circ$ spatial resolution granules were downloaded from AIRS website to obtain the desired output. Atmospheric parameters measured to include AST, CO, CH_4, SSKT, AT, RH, MSP $H2O_{vapour}$ and columnar O_3. The columnar O_3 data were also acquired from MMD. The data were collected in monthly intervals in 2009 for Subang station. These *in situ* data were used for comparison with the data from the developed regression equations and observed data from AIRS, in order to check the efficiency and accuracy of the equation.

2.2. Method of analysis

The measurements of columnar ozone during NEM season were found to be significantly different from those of SWM season depend on weather conditions and topography. The columnar O_3 has an inverse relationship with rain and a positive relationship with temperature. The columnar O_3 has maximum value during April–June because of the highest temperature which dominates most of the area and the increasing number of sunny hours in this period to reach its highest rates in May 8.7 hours day^{-1}.

PCA is a method that reduces data set dimensionality by performing a covariance analysis between factors. Fundamentally, PCA maximises the correlation between the original variables and new uncorrelated variables that is mutually orthogonal. The eigen-technique is used for special cases of factor analysis, and it transforms the original set of inter-correlation variables into a new set with an equal number of independent uncorrelated variables, which are linear combinations of the original variables. The MRA can lead to difficulties when the independent variables are highly correlated (multicollinearity). PCA is functional for extenuating the problem of multicollinearity. The principal components (PCs) are classified in a decreasing order according to the percentage of the variance that they account for. Although the number of PCs equals the number of independent original variables, most of the variation in the data can be described by the first few PCs that can be used to represent the original observations. Most of the variation in the data is explained by the first PC, and the second mode fits the remaining variance and creates a component that is uncorrelated with the first and so on.

Because the new variables from the PCA optimise spatial patterns and remove possible complications caused by multicollinearity, they become ideal to use as predictors in a regression equation. One such application is to obtain a varimax rotation of the original variable associated with each of the first few CPs, which can be utilised as a predictor in the linear regression model. The varimax rotation ensures that each variable is maximally correlated with only one PC and has a near-zero value that is shared with the other components (Jolliffe, 2002). The PCA is a simple and non-parametric method of extracting relevant information from confusing data sets; therefore, it is used abundantly in all forms of geophysical measurements (Hassanzadeh et al., 2008; Statheropoulos et al., 1998; Tian et al., 1989; Abdul-Wahaba et al., 2005).

In this study, we combine the multiple regression method and PCA to obtain regression equations for O_3 columnar value with other measured ambient atmosphere parameters as predictor variables. PCA was used to filter the data so that only the significant independent variables responsible for the ozone concentrations observed could be determined. The study concentrated on identification of factors that regulate ozone levels during NEM and SWM seasons. To better assess the measurement of the pair-wise associations among the variables, a correlation matrix was obtained for each data set.

The results of the PCA were used for principal component regression analysis, and a stepwise regression option was applied to select which PCs to enter the regression equation with O_3 as the dependent variable. The high loading variables were selected on the rotated principal components that were then used for inclusion in the ultimate regression model (Abdul-Wahaba et al., 2005;

Al-Alawi et al., 2008). The analysis of the data was carried out using the SPSS software, version 17.0.

3. Results and discussion

The aim of this section is to generate the regression equations for calculating the total column of ozone (O_3) in peninsular Malaysia from the standard monthly AIRX3STM products from six years' (2003–2008) data. Data for eight atmospheric parameters (AST, CO, CH_4, H_2O_{vapour}, SSKT, AT, RH, and MSP) were utilised to generate the regression equations of O_3 using both multiple linear and principal component regression methods. For better accuracy in developing the regression equations to measure O_3 the period was separated into two seasons (NEM and SWM seasons) because of the different impacts of both the NEM and SWM on the atmospheric parameters (AST, CO, CH_4, H_2O_{vapour}, SSKT, AT, RH, and MSP), in terms of the effects on climate or the amount of pollutants; they bring to the study area. Separate analyses were carried out for the NEM and SWM seasons. For both periods, the parameters were highly correlated.

3.1. Analysis of ozone data

Table 1a and b shows the Pearson correlation matrices of the variables for data from the NEM and SWM seasons for the study period. The columnar O_3 was negatively correlated with CH_4, H_2O_{vapour}, RH, and MSP, whereas it was positively correlated with CO, AST, SSKT, and AT during both the NEM and SWM season periods. These results from last three parameters (AST, SSKT, and AT) were expected because they are known as precursors of O_3. The increase in columnar O_3 concentration is associated with an increase in the values of AST, SSKT, AT, and CO and with a drop in the levels of CH_4, H_2O_{vapour}, RH, and MSP.

The cross correlation between the different variables is shown in Table 1a and b. During the NEM season, the columnar O_3 was particularly negatively correlated with RH (−0.844), SSKT (0.721), and AT (0.791) and negatively but weakly correlated with CH_4 (−0.353), H_2O_{vapour} (−0.351), and MSP (−0.334). In the SWM season, the columnar O_3 was highly and positively correlated with SSKT (0.704) and AT (0.632) and negatively correlated with RH (−0.808) and MSP (−0.627). In addition, columnar O_3 was negatively but weakly correlated with CH_4 and H_2O_{vapour}, and it was positively correlated with CO. The largest in magnitude among the latter were those with CH_4 (−0.397), followed by H_2O_{vapour} (−0.381).

3.2. Principal component analysis (PCA)

As explained in Section 2.2, the central idea of PCA is to decrease the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the difference present in the data set. The predictor variables were first transformed into an equal number of principal components (PCs). The objective was to get a small number of components that can explain most (60–90%) of the total variation in the predictor variables. After using the PCA, we maximise the loading of a predictor variable on one component by using varimax rotation. Generally, implementation of the PCA procedures followed by an orthogonal rotation method (varimax rotation) produces a ranked series of factors (Jolliffe, 2002). The results of the varimax rotation on the eight principal components together with the amount of variance accounted for by each component for the NEM and SWM seasons is summarised in Table 2a and b.

For a higher loading of a certain variable, the variation of other variables caused on a certain day by a particular principal component is greater. An eigenvalue greater than or equal to one for a principal component is commonly considered to be a criterion for statistical significance. As seen from Table 2a and b, the first three principal components accounted for more than 87% and 84% of the total variation for the NEM and SWM seasons, respectively.

For the NEM season, the first PC accounted for 48% of the total variation in the data. It loaded heavily on [RH], [MSP], [AT], [CO], and [H_2O_{vapour}]. The second PC, which accounted for approximately 27% of the total variation, was loaded heavily on [AST] and [SSKT] with a slight contribution from [RH], [AT], [CH_4], and [H_2O_{vapour}]. Principal components three and four loaded on [CH_4] and [H_2O_{vapour}], respectively. The rest of the principal components represented by the remaining variables accounted progressively for less of the total variation. The pattern for the SWM season data was only slightly different from that of the NEM season for the first three principal components (Table 2b). A multiple stepwise regression analysis of the principal component scores was then used to determine the original independent variables that are significant for the variation of the transformed columnar O_3 observation.

Table 1
Pearson correlation matrix of different variables for NEM and SWM season.

	AST	CO	O_3	CH_4	H_2O_{vapour}	SSKT	AT	RH	MSP
(a) NEM season									
AST	1	0.461	0.493	0.394	0.573	0.929	0.358	0.436	0.386
CO		1	0.209	−0.249	−0.385	−0.448	0.552	−0.672	−0.725
O_3			1	−0.353	−0.351	0.721	0.791	−0.844	−0.334
CH_4				1	0.234	0.363	0.303	0.224	−0.09
H_2O_{vapour}					1	0.516	−0.492	0.800	0.724
SSKT						1	0.322	−0.417	0.373
AT							1	−0.66	−0.89
RH								1	0.692
MSP									1
(b) SWM season									
AST	1	0.377	0.436	−0.105	0.682	0.755	0.313	0.534	0.383
CO		1	0.105	−0.331	−0.391	−0.27	0.368	−0.849	−0.499
O_3			1	−0.397	−0.381	0.704	0.632	−0.808	−0.627
CH_4				1	0.367	−0.122	−0.069	0.576	0.213
H_2O_{vapour}					1	0.307	−0.444	0.95	0.699
SSKT						1	−0.275	−0.424	0.301
AT							1	−0.538	−0.894
RH								1	0.735
MSP									1

Table 2
Rotated principal components loadings for NEM and SWM season.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
(a) NEM season								
AST	0.028	0.952	0.179	−0.106	0.118	−0.141	0.06	0.092
CO	−0.768	−0.141	0.454	−0.284	0.31	0.083	−0.02	−0.001
CH_4	0.079	0.224	0.635	0.164	−0.204	0.019	−0.005	0.003
H_2O_{vapour}	0.847	0.323	0.027	0.327	0.15	0.184	−0.096	0.066
SSKT	−0.091	0.976	0.089	−0.017	0.039	0.029	−0.012	−0.168
AT	−0.839	0.311	0.015	0.41	0.047	−0.134	0.053	0.032
RH	0.877	−0.383	0.076	0.154	0.123	0.097	0.169	−0.04
MSP	0.858	0.161	0.201	−0.461	−0.064	−0.052	0.015	0.034
Eigen value	4.304	2.467	1.069	0.646	0.218	0.183	0.049	0.048
% of Variance	47.82	27.41	11.87	7.178	2.42	2.03	0.548	0.529
Cumulative %	47.82	75.24	87.11	94.29	96.71	98.74	99.29	99.81
(b) SWM season								
AST	0.171	0.214	0.722	−0.057	0.085	0.192	0.061	0.013
CO	−0.355	−0.83	−0.103	−0.035	−0.303	−0.013	0.069	0.036
CH_4	−0.038	0.087	0.821	−0.029	0.95	0.003	0.013	0.008
H_2O_{vapour}	0.742	0.57	0.057	−0.528	0.189	0.094	0.294	0.014
SSKT	−0.265	0.924	−0.072	0.355	0.049	−0.314	−0.073	−0.02
AT	−0.861	−0.05	−0.02	0.203	0.012	0.045	−0.046	0.071
RH	0.875	0.558	−0.223	−0.491	0.201	0.026	0.1	0.181
MSP	0.813	0.249	0.448	−0.223	0.097	0.109	−0.039	0.155
Eigen value	4.307	2.12	1.165	0.709	0.438	0.119	0.069	0.044
% of Variance	47.85	23.56	12.94	7.87	4.86	1.323	0.771	0.488
Cumulative %	47.85	71.41	84.35	92.23	97.10	98.42	99.19	99.68

Table 3a and b summarises the results of the analysis for the NEM and SWM seasons, respectively.

3.3. Model fitting

Selecting a subset of the predictor variables that derive the columnar O_3 regression equation by using the multiple regression method was the main objective of this last section. The selected original independent variables with high loading associated with each principal component are included in a regression equation that had high coefficients of determination.

For both the NEM and SWM season observations, Tables 2 and 3 were used to match a PC in the regression analysis to independent variables. [RH] was selected from PC1, [SSKT] from PC2 and [CH_4] from PC3.

These three variables were then used as predictor variables in a subsequent regression analysis. The following regression equations for columnar value of O_3 [(PCA) O_3] were derived.

For the NEM season:

$$(\text{PCA1})O_3 = 201.128 + 0.203[\text{SSKT}] - 0.077[CH_4] - 0.425[\text{RH}] \tag{1}$$

For the SWM season:

$$(\text{PCA2})O_3 = 208.257 + 0.225[\text{SSKT}] - 0.095[CH_4] - 0.409[\text{RH}] \tag{2}$$

Table 3
Linear regression model for prediction of columnar O_3 using the principle components.

Predictors	Constant	PC1	PC2	PC3
(a) NEM season				
Adjusted *R* square		0.631	0.85	0.861
Estimated regression coefficient	201.128	0.794	0.922	0.931
(b) SWM season				
Adjusted *R* square		0.565	0.683	0.874
Estimated regression coefficient	208.257	0.755	0.831	0.935

3.4. Comparison and validation of the regression equations

To test the two columnar values of O_3 regression equations according to Eqs. (1) and (2), we checked the validation of both the (PCA1) O_3 NEM season and (PCA2) O_3 SWM season regression equations against the observed columnar value of O_3 obtained from AIRS. The validation was conducted with a linear regression correlation throughout the year 2009 over Subang station, in addition for two selected months, over Peninsular Malaysia, each from the NEM and SWM seasons: December and January and July and August 2009, respectively, as presented in Fig. 2.

Both NEM and SWM seasons model, using the three variables, [SSKT], [CH4], and [RH], yielded the strong correlation coefficient (R) represented by the values 0.845, 0.876, 0.916, 0.857 and 0.942 for December, January, July, August and Subang station for 2009, respectively. In addition, the asset values of adjusted coefficients (R^2) were 0.713, 0.767, 0.839, 0.735 and 0.887 for December, January, July, August and Subang station, respectively. The coefficients of the regressions were all statistically highly significant and ρ-values for all coefficients were less than 0.005 ($\rho < 0.005$). As for the independent samples *t*-test, SPSS displays relevant descriptive statistics, the (ρ-value) associated with the test statistic, and a confidence interval for the mean difference.

This strong high positively correlation shows a good indication of efficiency and accuracy of the both predicted NEM and SWM seasons regression equations. These results indicate that the predicted ozone from PCA is nearly the same as the observed ozone. The column of O_3 concentrations predicted by Eqs. (1) and (2) was plotted against the observed values from AIRS, and the results are presented in Fig. 2. The points tended to cluster along the 45° tangent, providing another indication of the model efficiency.

A slight variation and close fit, especially during the NEM season, is plainly evident in the comparison between predicted values of columnar O_3 with the mean observed and *in situ* measurements, plotted against months over Subang station in 2009, as presented in Fig. 3. The amplitude of the columnar ozone seasonal cycle generally increased to a natural peak in the late NEM season and early SWM season between mid-March and June, and there were disparities between the predicted and the observed values

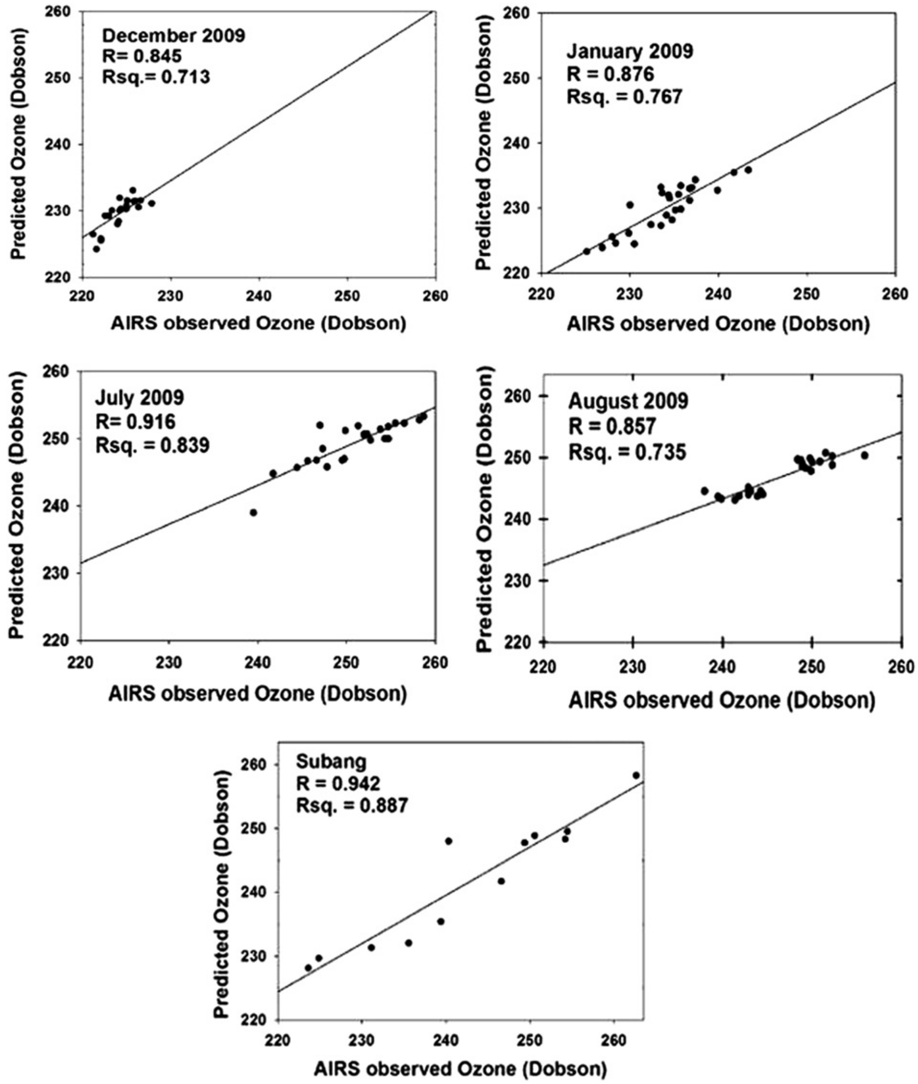

Fig. 2. Predicted vs. observed values of columnar O_3 from AIRS for the months of December, January, July, and August, and at Subang station for 2009.

resulting from increased temperatures and number of hours of sunshine during this period, especially in April and May, when temperature and hours of sunshine were at their maximum (Tangang et al., 2007).

The *in situ* measurement curve has ripples and fluctuations that are parallel to the predicted curve throughout the year. There was a small phase delay between the observed and *in situ* columnar ozone values, which was almost the same as that between *in situ* and predicted values. The closed compatibility in Fig. 3 between predicted values of columnar O_3 with the mean observed and *in situ* measurements in both dramatic drop and increase in columnar O_3 in Subang station shows a good indication of efficiency and

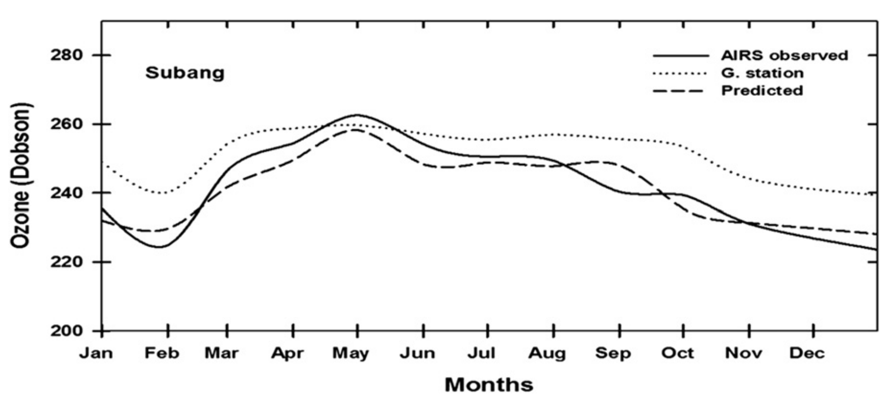

Fig. 3. Observed ozone from AIRS (solid line), *in situ* measurements (dotted line), and predicted value (dashed line) in 2009 for the Subang station.

accuracy of the both predicted PCA1 and PCA2 regression equations.

4. Conclusions

It is known that ozone is recognized as one of the main pollutants that degrade air quality, and the stratospheric ozone that is significant. The objective of this study was to obtain accurate prediction models (i.e., models that depend on as few variables as necessary) for columnar ozone with other atmosphere parameter's data as predictor variables.

Six-year (2003–2008) satellite data were employed to develop the regression equations for calculating columnar O_3 for NEM and SWM seasons over peninsular Malaysia using both multiple linear and principal component regression methods. In addition to analyse the effects of the atmosphere parameters on ozone column value over study area.

Ozone concentrations were negatively correlated with CH_4, H_2O_{vapour}, RH, and MSP, whereas it was positively correlated with CO, AST, SSKT, and AT during both the NEM and SWM season periods. These results from last three parameters (AST, SSKT, and AT) were expected because they are known as precursors of columnar O_3. The increase in columnar O_3 concentration is associated with an increase in the values of AST, SSKT, AT, and CO and with a drop in the levels of CH_4, H_2O_{vapour}, RH, and MSP.

The result of fitting the best regression equations for the columnar O_3 data using eight of the independent variables gave about the same values of the R ($\approx$0.93) and R^2 ($\approx$0.86) for both the NEM and SWM seasons. The common variables that appeared in both regression equations were SSKT, CH_4 and RH, and the principal precursor of the columnar O_3 value in both the NEM and SWM seasons was SSKT.

The ability to evaluate the columnar O_3 regression equation was demonstrated through comparisons with the observed columnar O_3 from AIRS and *in situ* data from Subang station in 2009. The results reveal a close agreement between the predicted data and the AIRS observed data for columnar O_3 values throughout the year. Similar features appeared between the predicted columnar O_3 and *in situ* measurements. Validation using linear regression correlation was done for columnar O_3 with observed O_3 from AIRS. The results revealed that columnar O_3 had values nearly same as the observed columnar O_3 from AIRS. In addition, columnar O_3 had good accuracy and efficiency in all cases resulting from good correlation coefficients (R, 0.845–0.942) and adjusted coefficients (R^2, 0.713–0.887).

Overall, these results clearly indicate the advantage of using satellite AIRS data and correlation analysis to investigate the impact of atmosphere parameters on columnar O_3 over peninsular Malaysia. The validation and comparison conducted in this study successfully demonstrate the high accuracy of the regression equation, supporting the research. The study shows that AIRS data are useful and suitable to be utilized to investigate the impact of atmosphere parameters in Southeast Asia. The present study involved the western part of Malaysia, where the models have been generated for this region.

Acknowledgements

The authors gratefully acknowledge the National Aeronautics and Space Administration (NASA) Goddard Earth Sciences Data Information and Services Centre (DISC) for the provision of the AIRS data and images used in this paper. We would like to thank the technical staff who participated in this project. Thanks are also extended to USM for support and encouragement.

References

Abdul-Wahab, S.A., 2001. IER photochemical smog evaluation and forecasting of short-term ozone pollution levels with artificial neural networks. Trans. ICHEME Process. Saf. Environ. Prot. 79, 117–128.

Abdul-Wahaba, S.A., Bakheitb, C.S., Al-Alawi, Saleh M., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ. Model. Softw. 20, 1263–1271.

Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S., 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environ. Model. Softw. 23, 396–403.

Bian, J., Gettelman, A., Chen, H., Pan, L.L., 2007. Validation of satellite ozone profile retrievals using Beijing ozonesonde data. J. Geophys. Res. 112, D06305.

Blankinship, D.J., 1996. A Discussion of the Spatial and Temporal Variability of Ozone Concentrations along the Front Range of Colorado. University of Colorado at Boulder (Program in Atmospheric and Oceanic Sciences).

Bossioli, E., Tombrou, M., Dandou, A., Soulakellis, N., 2007. Simulation of the effects of critical factors on ozone formation and accumulation in the greater Athens area. J. Geophys. Res. 112, D02309–D02319.

Clerbaux, C., Hadgi-Lazaro, J., Turquety, S., Metie, G., Coheur, F.P., 2003. Trace gas measurements from infrared satellite for chemistry and climate applications. Atmos. Chem. Phys. 3, 1495–1508.

Dasimah, Bt Omar, 2009. Urban form and sustainability of a hot humid city of Kuala Lumpur. Eur. J. Soc. Sci. 8, 353–359.

Department Of Environment (DOE), 2004. Malaysia Environmental Quality Report. Ministry of Natural Resources and Environment Malaysia, Petaling Jaya, Sasyaz Holdings Sdn Bhd.

Dousset, B., Gourmelon, F., 2003. Satellite multi-sensor data analysis of urban surface temperatures and landcover. ISPRS J. Photogramm. Remote Sens. 58, 43–54.

Duenas, C., Fernandez, M.C., Canete, S., Carretero, J., Liger, E., 2002. Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. Sci. Total Environ. 299, 97–113.

Fishbein, E., Hearty, T., Lee, S.-Y., Irion, F.W., Kahn, B., Manning, E., Blaisdell, J., Susskind, J., Iredell, L., Barnet, C., Maddy, E., Rosenkranz, P., McMillan, W.W., Machado, S.D., Knuteson, R., 2007. AIRS Version 5 Release Level 2 Standard Product QuickStart. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

Freijer, J.I., van Eijkeren, J.C.H., van Bree, L., 2002. A model for the effect on health of repeated exposure to ozone. Environ. Model. Softw. 17, 553–562.

Hassanzadeh, S., Hosseinibalam, F., Omidvari, M., 2008. Statistical methods and regression analysis of stratospheric ozone and meteorological variables in Isfahan. Phys. A 387, 2317–2327.

Illingworth, S.M., Remedios, J.J., Boesch, H., Ho, S.-P., Edwards, D.P., Palmer, P.I., Gonzi, S., 2011. A comparison of OEM CO retrievals from the IASI and MOPITT instruments. Atmos. Meas. Tech. 4, 775–793.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer, New York.

Maenhaut, Willy, Cornille, Philippe, Pacyna, Jozef M., Vitols, Val, 1989. Trace element composition and origin of the atmospheric aerosol in the Norwegian arctic. Atmos. Environ. 23 (11), 2345–2637.

McAdams, H.T., Crawford, R.W., Hadder, G.R., 2000. A Vector Approach to Regression Analysis and Its Application to Heavy-duty Diesel Emissions. Society of Automotive Engineers, Inc. Contract with the Energy Division of Oak Ridge National Laboratory (ORNL), Contract No. DE-AC05-00OR22725.

Mintzer, M.I., Miller, S.A., 1987. The ozone layer: Its protection depends on international cooperation. Environ. Sci. Technol. 21 (12), 1167–1169.

Morris, A.G., Hersey, S., Thompson, M.A., Pawson, S., Nielsen, E.J., Colarco, R.P., Mcmillan, W.W., Stohl, A., Turquety, S., Warner, J., Johnson, J.B., Kucsera, L.T., Larko, E.D., Oltmans, G.S., Witte, C.J., 2006. Alaskan and Canadian forest fires exacerbate ozone pollution over Houston, Texas, on 19 and 20 July 2004. J. Geophys. Res. 111, D24S03.

National academy press (NAP), 1991. Rethinking the Ozone Problem in Urban and Regional AirPollution, Washington, D.C., pp. 93–95.

Pagano, T.S., Chahine, M.T., Aumann, H.H., Tian, B., Lee, S.Y., Olsen, E., Lambrigtsen, B., Fetzer, E., Irion, F.W., McMillan, W., Strow, L., Fu, X., Barnet, C., Goldberg, M., Susskind, J., Blaisdell, J., 2006. Remote sensing of atmospheric climate parameters from the atmospheric infrared sounder. IEEE Trans. Geosci. Remote Sens., 2386–2389.

Pittman, J.V., Pan, L.L., Wei, J.C., Irion, F.W., Liu, X., Maddy, E.S., Barnet, C.D., Chance, K., Gao, R.S., 2009. Evaluation of AIRS, IASI, and OMI ozone profile retrievals in the extratropical tropopause region using in situ aircraft measurements. J. Geophys. Res. 114, D24109.

Rajab, J.M., MatJafri, M.Z., Tan, F., Lim, H.S., Abdullah, K., 2011. Analysis of Ozone column burden in Peninsular Malaysia retrieved from Atmosphere Infrared Sounder (AIRS) data: 2003–2009. IEEE International Conference on Imaging Systems and Techniques (IST), 978-1-61284-896-9/11/$26.00 ©2011 IEEE.

Rajab, J.M., Mat Jafri, M.Z., Lim, H.S., Abdullah, K., 2012. Regression analysis in modelling of air surface temperature and factors affecting its value in Peninsular Malaysia. Opt. Eng. 51 (10), 101702.

Sebald, L., Treffeisen, R., Reimer, E., Hies, T., 2000. Spectral analysis of air pollutants. Part 2: Ozone time series. Atmos. Environ. 34, 3503–3509.

Statheropoulos, M., Vassiliadis, N., Pappa, A., 1998. Principal component and canonical correlation analysis for examining air pollution and meteorological data. Atmos. Environ. 32, 1087–1095.

Strow, L.L., Hannon, S.E., Weiler, M., Overoye, K., Gaiser, S.L., Aumann, H.H., 2003. Prelaunch spectral calibration of the atmospheric infrared sounder (AIRS). IEEE Trans. Geosci. Remote Sens. 41, 274–286.

Struijs, J., van, Dijk A., Slaper, H., vav Wijnen, H.J., Velders, G.J.M., Chaplin, G., Huijbregts, M.A.J., 2010. Spatial- and time-explicit human damage modeling of ozone depleting substances in life cycle impact assessment. Environ. Sci. Technol. 44 (1), 204–209.

Suhaila, J., Jemain, A.A., 2007. Fitting daily rainfall amount in Malaysia using the normal transform distribution. J. Appl. Sci. 7, 1880–1886.

Suhaila, J., Jemain, A.A., 2009. Investigating the impacts of adjoining wet days on the distribution of daily rainfall amounts in Peninsular Malaysia. J. Hydrol. 368, 17–25.

Tangang, F.T., Juneng, L., Ahmad, S., 2007. Trend and interannual variability of temperature in Malaysia: 1961–2002. Theor. Appl. Climatol. 89, 127–141.

Tian, Y.I., Biswas, P., Pratsinis, S.E., Hsieh, W.M., 1989. Principal component analysis for particulate source resolution in cleanrooms. November/December. J. Environ. Sci., 22–27.

Tiwari, Y.K., Gloor, M., Engelen, R., Rodenbeck, C., Heimann, M., 2005. Comparing model predicted atmospheric CO_2 with satellite retrievals and in-situ observations – implications for the use of upcoming satellite data in atmospheric inversions. Geophys. Res. Abstr. 7, A-09823.

Vaidya, O.C., Howell, G.D., Leger, D.A., 2000. Evaluation of the distribution of mercury in Lakes in Nova Scotia and Newfoundland (Canada). Water Air Soil Pollut. 117 (1–4), 353–369.

Wang, H., Zou, X., Li, G., 2012. An improved quality control for AIRS total column ozone observations within and around Hurricanes. J. Atmos. Oceanic Technol. 29, 417–432.

Wong, C.L., Venneker, R., Uhlenbrook, S., Jamil, A.B.M., Zhou, Y., 2009. Variability of rainfall in Peninsular Malaysia. Hydrol. Earth Syst. Sci. Discuss. 6, 5471–5503.

Ye, H., Fetzer, E.J., Bromwich, D.H., Fishbein, E.F., Olsen, E.T., Granger, S.L., Lee, S.Y., Chen, L., Lambrigtsen, B.H., 2007. Atmospheric total precipitable water from AIRS and ECMWF during Antarctic summer. Geophys. Res. Lett. 34, L19701.

Zhang, L., Li, Q.B., Murray, L.T., Luo, M., Liu, H., Jiang, J.H., Mao, Y., Chen, D., Gao, M., Livesey, N., 2012. A tropospheric ozone maximum over the equatorial Southern Indian Ocean. Atmos. Chem. Phys. 12 (9), 4279–4296.