Front. Earth Sci. 2016, 10(1): 63–73
DOI 10.1007/s11707-015-0521-8

RESEARCH ARTICLE

A statistical model to predict total column ozone in Peninsular Malaysia

K. C. TAN (✉), H. S. LIM, M. Z. MAT JAFRI

School of Physics, Universiti Sains Malaysia, Penang 11800, Malaysia

Abstract This study aims to predict monthly columnar ozone in Peninsular Malaysia based on concentrations of several atmospheric gases. Data pertaining to five atmospheric gases (CO_2, O_3, CH_4, NO_2, and H_2O vapor) were retrieved by satellite scanning imaging absorption spectrometry for atmospheric chartography from 2003 to 2008 and used to develop a model to predict columnar ozone in Peninsular Malaysia. Analyses of the northeast monsoon (NEM) and the southwest monsoon (SWM) seasons were conducted separately. Based on the Pearson correlation matrices, columnar ozone was negatively correlated with H_2O vapor but positively correlated with CO_2 and NO_2 during both the NEM and SWM seasons from 2003 to 2008. This result was expected because NO_2 is a precursor of ozone. Therefore, an increase in columnar ozone concentration is associated with an increase in NO_2 but a decrease in H_2O vapor. In the NEM season, columnar ozone was negatively correlated with H_2O (-0.847), NO_2 (0.754), and CO_2 (0.477); columnar ozone was also negatively but weakly correlated with CH_4 (-0.035). In the SWM season, columnar ozone was highly positively correlated with NO_2 (0.855), CO_2 (0.572), and CH_4 (0.321) and also highly negatively correlated with H_2O (-0.832). Both multiple regression and principal component analyses were used to predict the columnar ozone value in Peninsular Malaysia. We obtained the best-fitting regression equations for the columnar ozone data using four independent variables. Our results show approximately the same R value (≈ 0.83) for both the NEM and SWM seasons.

Keywords ozone, SCIAMACHY, principal component analysis, Peninsular Malaysia

Received July 26, 2014; accepted March 2, 2015

E-mail: kctan@usm.my

1 Introduction

While ozone is one of the least prevalent gases present in our atmosphere, it is an important chemical constituent that is implicated in the atmospheric energy budget as well as atmospheric chemistry, air quality, and global change (Dueñas et al., 2004; Ahammed et al., 2006; Lin et al., 2008). Even though ozone is recognized as one of the main greenhouse gases and exerts significant effects on the radiation budget of the atmosphere, it is considered to be a secondary pollutant (Wu and Chan, 2001). Abrupt changes in atmospheric ozone caused by high levels of surface ozone and anthropogenic emissions may create environmental problems and contribute to climate change (Vingarzan, 2004).

Ozone consists of critical atmospheric trace gases in the stratosphere and troposphere because it functions as an oxidant and greenhouse gas (Toh et al., 2013). The majority of ozone gases are found in the upper part of the atmosphere, particularly the stratosphere, more than 10 km above the Earth's surface. Approximately 90% of atmospheric ozone is contained in the ozone layer of the stratosphere while the remaining 10% is found in the troposphere. The troposphere extends from the Earth's surface to the tropopause, from 10 km to 18 km, and is the atmospheric layer in which humans live and emit chemical compounds from anthropogenic activities (Reddy et al., 2012).

On the ground, ozone is a harmful pollutant that endangers lung tissues and the ecosystem (Bian et al., 2007). In the presence of sunlight, nitrogen oxides (NO_x) react with volatile organic compounds (VOCs) to form ozone. VOCs are emitted from various sources, such as motor vehicles and other industrial sources. On the other hand, stratospheric ozone is beneficial because it is the only atmospheric component that absorbs harmful ultraviolet (UV) rays from the sun before they can reach the Earth's surface. Without stratospheric ozone, humans would be exposed to large amounts of ultraviolet-B (UV-

B) radiation with wavelengths ranging from 0.2 μm to 0.28 μm (Tan et al., 2012a, 2014b).

Southeast Asia has recently experienced rapid economic and industrial development that has resulted in an increase in the amount of air pollutants released into the atmosphere. These pollutants have created significant issues with tropospheric ozone chemistry that need to be addressed. In Southeast Asia and other tropical countries, biomass burning, particularly forest fires, also contributes to the amount of tropospheric ozone (Pochanart et al., 2001; Tan et al., 2012b). Although numerous studies on ozone in Southeast Asia have been conducted, there have been few that focus specifically on the ozone trends over Peninsular Malaysia even though Malaysia has undergone rapid economic development and urbanization in recent years, resulting in an increase in the consumption of fossil fuels with the concomitant increase in emissions of air pollutants, particularly in industrial areas and cities (Tan et al., 2012c).

In Malaysia, previous studies have relied on ground station data due to the lack of observational greenhouse gas data since studies using satellite data have not considered equatorial areas (Tan et al., 2014a). In the last three years, however, there have been many studies employing greenhouse gas satellite data from Malaysia and equatorial areas. These studies are especially important because they can account for the influence of the monsoon on atmospheric parameters. The Malaysian climate is dominated by a strong northeast monsoon (NEM) from November to April and a southwest monsoon (SWM) from May to October. These monsoons exert different influences on atmospheric parameters in Malaysia depending on the climate or the amount of pollutants as well as the contribution of the many regional sources of pollutants (Rajab et al., 2013). Thus, satellite data can be useful in investigating the relationship between atmospheric variables and atmospheric ozone over Peninsular Malaysia.

Satellite remote sensing is one of the most effective approaches for monitoring the distributions of greenhouse gases on a global scale at very high spatial and temporal resolution (Baker et al., 2010), and provides an alternative for evaluating the influence of human anthropogenic activity on climate change. The free, downloadable data from the satellite scanning imaging absorption spectrometer for atmospheric chartography (SCIAMACHY) onboard the ENVISAT can be used to observe the Earth's greenhouse gas concentrations (Tan et al., 2014b).

Multiple regression analysis (MRA) is one of the most frequently used methodologies to determine the dependence of a response variable on several independent variables and is usually applied to obtain a linear input-output model for a specific data set (Al-Alawi et al., 2008). However, the regression approach can encounter major problems when independent variables are correlated with each other (Abdul-Wahab et al., 2005). Hence, an alternative method should be used to eliminate multicollinearity; multivariate data analysis (MDA), in particular, is an effective alternative technique used to address this limitation. MDA techniques have been widely applied in environmental research, specifically in trend and relationship analysis (Statheropoulos et al., 1998). Many MDA methods can be used, but principal component analysis (PCA) is one of the most common techniques used in air quality studies to analyze large environmental data sets (Vaidya et al., 2000). In previous studies, PCA methods have been successfully used to identify important factors influencing ozone concentrations and to examine ozone variations (Lengyel et al., 2004).

Primarily, this study aimed to develop a regression model of the columnar ozone over Peninsular Malaysia during the NEM and SWM seasons, using as predictors concentrations of four atmospheric gases (carbon dioxide: CO_2, methane: CH_4, water vapor: H_2O vapor, and nitrogen dioxide: NO_2). Seven years of satellite data, from January 2003 to December 2009, were considered in this study. Data from six of these years (January 2003 to December 2008) were acquired to analyze and develop predictive regression models of columnar ozone, and the data from 2009 were used to validate and compare our results. The satellite data were analyzed in terms of atmospheric parameters (O_3, CO_2, CH_4, H_2O vapor, and NO_2) obtained from SCIAMACHY nadir spectra levels 2 and 3. This is the first study to use SCIAMACHY satellite data to model ozone in Peninsular Malaysia.

2 The SCIAMACHY instrument

The SCIAMACHY, which is on board the ENVISAT as part of the atmospheric chemistry payload of the European Space Agency's third Earth observation mission, is the first satellite instrument whose measurements are sufficiently precise and sensitive to detect all greenhouse gases. As such, observations are possible at all altitudes down to the Earth's surface (Schneising et al., 2008a). The ENVISAT flies at an altitude of 795 km above the Earth's surface in near synchronous polar orbit and passes by the equator with a descending node at 10:00 a.m. local time.

The SCIAMACHY is a passive remote sensing instrument that measures reflected, scattered, and transmitted solar radiation from the atmosphere. It is comprised of eight spectral channels between 214 and 2,380 nm at a moderate spectral resolution between 0.2 nm and 1.4 nm (Bracher et al., 2005). The spectral region from 214 nm to 1,750 nm is determined in six adjacent channels, and the two remaining channels cover the regions from 1,940 nm to 2,040 nm and 2,265 nm to 2,380 nm, respectively. An extraordinary characteristic of the SCIAMACHY is that its spectroscopic observations are based on alternating nadir and limb viewing geometries. For the total columnar ozone

(in DU) retrieved from the SCIAMACHY, the effective spatial resolution varies between 30 km along the track and between 30 km and 240 km across the track. The Weighting Function Modified-Differential Optical Absorption Spectroscopy (WFM-DOAS) algorithm was developed to retrieve atmospheric data at the Institute of Environmental Physics of the University of Bremen in Germany.

3 Data and methodology

3.1 Site description and data collection

Peninsular Malaysia is located between 1° to 7° north latitude and 99° to 105° east longitude in Southeast Asia (south of Thailand, north of Singapore, and east of the Indonesian island of Sumatra). The area of Peninsular Malaysia is approximately 131,587 km^2 with an estimated population of 21 million (Fig. 1) (Tan et al., 2014b).

Peninsular Malaysia experiences a tropical climate throughout the year; the weather is warm and humid with temperatures ranging from 20°C to 32°C (Omar, 2009). The climate in this region is considerably influenced by the mountainous topography and complex land-sea interactions. Intra-seasonal and intra-decadal fluctuations, such as the El Niño-Southern Oscillation, Indian Ocean Dipole, and Madden Julian Oscillation, significantly influence the inter-annual climatic variability of Malaysia (Tan et al., 2014a). The highest monthly average temperatures occur in April, May, July, and August whereas the lowest average monthly temperatures are recorded from November to January.

The two rainy seasons in Peninsular Malaysia are due to the effect of the NEM from November to March and the SWM from May to September. Two inter-monsoon seasons occur in between (Yonemura et al., 2002). The SWM is drier than the NEM, which brings higher amounts of rainfall to the country. Lighting and variable wind thunderstorms develop in the afternoon during the inter-monsoon periods (www.met.gov.my).

3.2 Data acquisition

The available temporal resolutions of the SCIAMACHY standard products are daily, every 6 days, and monthly (Richter et al., 2005). In this study, all of the retrieved data (CO_2, CH_4, H_2O vapor, NO_2, and O_3) from the SCIAMACHY satellite were acquired and developed by the Institute of Environmental Physics (IUP) at the University of Bremen, Germany. Both of the SCIAMACHY products acquired from WFM-DOAS (WFM-DOAS version 2.1 Level 3 of XCO_2 and WFM-DOAS version 2.0.2 Level 3 of XCH_4) were developed by the Institute of Environmental Physics (IUP) at the University of Bremen

Fig. 1 Geographical features of the study area (Source: http://www.mapsofworld.com/lat_long/malaysia-lat-long.html and http://en.wikipedia.org/wiki/Peninsular _Malaysia).

(Schneising et al., 2008b). XCO_2 is the dry air column averaged mole fraction of CO_2 in ppm (parts per million) while XCH_4 is the dry air column averaged mole fraction of CH_4 in ppb (parts per billion). The SCIAMACHY WFM-DOAS Level 3 product contains gridded data at a monthly resolution that are produced from the corresponding Level 2b files filtered according to the recommended approach for each gas. The XCO_2 and XCH_4 files are located in Level3_XCO2_monthly_grid_QUALgood and Level3_XCH4_monthly_grid_QUALgood, respectively.

The scientific algorithm, WFM-DOAS, was developed to retrieve CO_2 columns from the SCIAMACHY nadir spectra (Schneising et al., 2008b. WFM-DOAS was developed at the Institute of Environmental Physics (IUP) at the University of Bremen, Germany,). The WFM-DOAS is a method that simultaneously retrieves the NIR nadir measurements of the SCIAMACHY instrument in the CO_2 absorption band from the spectral region of 1,558 nm to 1,594 nm and the oxygen-A absorption band from the spectral region of 755 nm to 775 nm to generate the dry air column averaged XCO_2 (in ppmv). The WFM-DOAS was also used to simultaneously retrieve the near-infrared measurements of the SCIAMACHY instrument in the CO_2 absorption band from the 1558−1594 nm spectral region and the oxygen-A absorption band at the 755−775 nm spectral region to generate the dry air column averaged mixing ration of methane, XCH_4 (in ppbv) assuming a constant CO_2 mole fraction of 370 ppm (Buchwitz et al., 2005).

In addition, the SCIAMACHY daily H_2O total column data are retrieved from the spectral measurement of the visible wavelength region of approximately 700 nm using the Air Mass Corrected-DOAS (AMC-DOAS) method. The SCIAMACHY H_2O vapor level 2 Version 1.0 data are derived from the SCIAMACHY reprocessed level 1b V5 data set. The measurement can only be made on the daytime and cloud-free ground scenes because the AMC-DOAS method is used to analyze data in the visible spectral range (Mieruch et al., 2008). The extraordinary quality of the AMC-DOAS method is that the derived H_2O columns are not affected by the calibration using radiosonde data, which is frequently applied to data in the microwave spectral region. The AMC-DOAS algorithm is based on the DOAS approach, which is used to process the information found in the differential absorption structures.

The retrieval of the SCIAMACHY daily level 2 NO_2 tropospheric column version 0.7 is based on the DOAS method (Richter et al., 2004), which retrieves the slant column density (SCD) along the light path through the atmosphere for a given spectral window. Some factors affect the slant column, such as the solar zenith angle, the viewing geometry and the amount and vertical distribution of the absorber in the atmosphere. After accounting for these, the analysis focuses on the correction for stratospheric absorption, which can be performed by deducting the NO_2 column over a clean study area. Finally, an air mass factor (AMF) based on radiative transfer calculations is used to convert the balance of the tropospheric slant column to a geometry-independent tropospheric vertical column.

The implementation of the WFM-DOAS retrieval algorithm successfully improves the quality of total column O_3 version 1.0 level 2 data (daily), which is retrieved from between 325 nm and 335 nm at a spectral resolution of approximately 0.2 nm. The WFM-DOAS algorithm has been used to retrieve total column O_3 SCIAMACHY nadir spectra (Stephens et al., 2007), and it can directly retrieve vertical column O_3 largely due to its vertically integrated O_3 weighting functions, which are then compared with O_3 cross-sections of sun-normalized radiances (Coldewey-Egbers et al., 2005). In addition, WFM-DOAS also considers slant column path length modulation as a wavelength function. The standard DOAS algorithm generally ignores this consideration.

All of the standard products from the SCIAMACHY nadir spectral levels 2 and 3 are downloaded from the SCIAMACHY website, and the following steps are involved in converting these data from the ASCII file into a table in MS-Excel. The data for each parameter are extracted and added to the same table. The parameters that were selected in this study are O_3, CO_2, CH_4, H_2O vapor, and NO_2. The observed CO_2 and CH_4 ($0.5°\times0.5°$), H_2O (non-gridded data), and NO_2 ($0.125°\times0.125°$) values were converted into $1°\times1.25°$ latitude/longitude projections to be at the same spatial resolution as the O_3 data. The converted (CO_2, CH_4, H_2O vapor, and NO_2) data in the new $1°\times1.25°$ spatial resolution were combined with the O_3 data, which is already at the $1°\times1.25°$ spatial resolution.

Previous research has demonstrated the influence of meteorological parameters, such as wind speed and wind direction, on ozone concentrations. This study examines the relationship between atmospheric variables and ozone. In addition, the SCIAMACHY does not provide meteorological parameter data, so it is impossible to incorporate the meteorological variables into the prediction models. Indeed, data from different satellites cannot be combined with the SCIAMACHY data due to the different spatial and temporal resolutions. Therefore, this study focuses on predicting columnar ozone using only atmospheric variables.

3.3 Method of analysis

The method most frequently used for the meteorological prediction of ozone is MRA (Tan et al., 2013). MRA is used to express the dependence of a response variable on several independent variables to obtain a linear input-output model for a given data set (Al-Alawi et al., 2008). PCA is a method that optimizes the correlation between a set of variables to form new variables that are uncorrelated or mutually orthogonal. This approach involves conduct-

ing an analysis of covariance between factors to reduce the dimensionality of environmental data sets. Compared with MRA, PCA performs well in terms of handling independent variables, but MRA cannot be applied when predictors are used in a regression equation, particularly in cases in which independent variables are highly correlated (multi-collinearity). Each principal component (PC) accounts for a percentage of the variance, which decreases as the order of each PC increases (Azid et al., 2014). In general, the number of independent variables is represented by the number of PCs. However, original observations are usually only attributed to the first few PCs in most of the variations in the data set. Thus, the first PC explains the majority of the variation in the data followed by the second mode that fits the remaining variance and so on.

PCA is suitable for regression equations because the new variables that arise eliminate problems associated with multi-collinearity, thus optimizing spatial patterns.. Predictors in linear regression analysis can be selected by obtaining a varimax rotation of the original variables associated with the first few PCs. The main purpose of performing varimax rotation is to ensure that one PC is maximally correlated with each variable and has a near zero value with the other components (Statheropoulos et al., 1998). Furthermore, PCA has been performed in all forms of geophysical measurements, particularly when confusing data are obtained, because this approach is simple and non-parametric (Vaidya et al., 2000).

In the current study, PCA and MRA were combined to establish a more accurate prediction model of the columnar ozone value using other atmospheric parameters as predictor variables. PCA is purposely used to filter large amounts of data to determine only the significant independent variables affecting the observed ozone levels. This study focused on the factors that affect ozone concentrations during the NEM and SWM seasons, and a correlation matrix was generated for each data set to assess the pairwise associations among the various variables. The results from the PCA were used in a stepwise principal component regression analysis. The selected PCs were used in the regression equation, in which ozone was the dependent variable. The selected variables with high loadings on the rotated PCs were then used in MRA.

4 Results and discussion

4.1 Analysis of ozone data

Table 1 shows the Pearson correlation matrices of the variables for the NEM and SWM seasons that occurred during the study period. Statistically significant correlation coefficients ($\rho < 0.05$) are highlighted in bold. The columnar ozone was negatively correlated with H_2O but positively correlated with CO_2 and NO_2 during both the NEM and SWM seasons from 2003 to 2008. This result was expected because NO_2 is a precursor of ozone. Therefore, an increase in the columnar ozone concentration is associated with an increase in NO_2 and a decrease in the H_2O level. In the NEM season, the columnar ozone was negatively correlated with H_2O (-0.847) but positively correlated with NO_2 (0.754) and CO_2 (0.477); the columnar ozone was also negatively but weakly correlated with CH_4 (-0.035). In the SWM season, the columnar ozone was highly positively correlated with NO_2 (0.855), CO_2 (0.572), and CH_4 (0.321). The columnar ozone was also highly negatively correlated with H_2O (-0.832).

H_2O vapor had a significant impact on the chemical and radiative characteristics of the stratosphere. Increasing amounts of H_2O vapor may give rise to radiative cooling of the stratosphere and affect chemical processes , and in this study, H_2O vapor induced changes in O_3 concentrations and caused radiative responses in the stratosphere. The highly reactive molecule known as hydroxyl (OH), which is produced by the photochemical breakdown of O_3 in the presence of H_2O vapor, will directly destroy O_3 in both the lower and upper stratosphere. OH can react with many pollutants, which in turn depletes O_3. Thus, increasing H_2O vapor gives rise to more radical OH and thus greater

Table 1 Pearson correlation matrices of the different variables in the NEM and SWM seasons

Season	Variable	CO_2	CH_4	O_3	H_2O	NO_2
NEM season	CO_2	1	0.366	**0.477**	0.316	0.215
	CH_4		1	−0.035	0.243	0.112
	O_3			1	**−0.847**	**0.754**
	H_2O				1	−0.371
	NO_2					1
SWM season	CO_2	1	0.39	**0.572**	−0.252	0.337
	CH_4		1	0.321	0.127	−0.161
	O_3			1	**−0.832**	**0.855**
	H_2O				1	−0.29
	NO_2					1

O_3 loss. Furthermore, Tian et al. (2009) found that the chemically induced effects of increasing H_2O vapor led to an overall decrease (approximately 1%) of O_3 in the tropics.

The results showed that columnar ozone correlated negatively with CH_4 in the NEM but positively in the SWM. During the NEM season, the production of CH_4 is low due to the fewer sunny hours. Tropical cyclones impact the weather of Peninsular Malaysia as well and significantly affect the emission of CH_4. Furthermore, the minimal hydroxyl (OH) levels reduce oxidation with CH_4 and the concentration of columnar ozone in the atmosphere because the primary CH_4 sink of is the oxidation of CH_4 by OH and the resulting formation of ozone (Johnson et al., 2002). Therefore, CH_4 fluxes were relatively low during the NEM season. During the SWM, the CH_4 concentration peak in the atmosphere was caused by the biogenic sources of CH_4, which usually include the fermentation of organic waste, swamps, and paddy rice fields (Wang et al., 2001). Due to urban expansion and the increasing population, waste is increasing as well. The higher seasonal temperature degrades organic substances to CH_4 at a faster rate and increases the concentration of CH_4 in the atmosphere. In addition, non-biogenic sources of CH_4, such as the combustion of fossil fuels, also increase the concentration of CH_4 in the atmosphere in Peninsular Malaysia.

4.2 PCA results

A significant degree of multi-collinearity was observed among the atmospheric parameters due to the strength of their correlations (Table 1). PCA can be used to effectively reduce the dimensionality of a data set that involves a large number of interrelated variables. We initially transformed the predictor variables into an equal number of PCs to obtain a small number of components that could explain the majority (approximately 60% to 90%) of the total variation in the predictor variables.

After transforming these variables, we maximized the loading of a predictor variable on one component using a varimax rotation; application of PCA followed by an orthogonal rotation method (varimax rotation) yielded a ranked series of factors. Table 2 summarize the results of the varimax rotation on the four PCs and the extent of the variance accounted for by each component in the NEM and SWM seasons. For each particular PC, the variation was attributed to the loading of a variable. In PCA, the loadings with absolute values $> 50\%$ are selected to interpret the PC (Abdul Wahab et al., 2005). An eigenvalue $\geqslant 1$ for a PC is one of the criteria indicating its statistical significance (Kaiser criterion).

The first two PCs accounted for $> 69\%$ and $> 68\%$ of the total variation in NEM and SWM seasons, respectively (Table 2). In the NEM season, the first PC accounted for 42% of the total variation in the data set, and this PC loaded heavily on $[H_2O]$ with a small contribution from $[CO_2]$. The second PC, which accounted for approximately 27% of the total variation, was loaded heavily on $[NO_2]$ with a slight contribution from $[CH_4]$. The remaining PCs were represented by the rest of the variables, which accounted for a small proportion of the total variation. The trend in the data for the SWM season was similar to that of the NEM season for the first two PCs (Table 2). As the main factors responsible for the first and second PCs, the atmospheric variables identified by the PCA were selected as independent variables and used in the stepwise multiple regression analysis; only the original independent variables that were significant in explaining the variation in the

Table 2 Rotated principal component loadings for the NEM and SWM seasons

Season	Variable	PC1	PC2	PC3	PC4
NEM season	CO_2	−0.537	−0.146	0.236	0.003
	CH_4	0.167	0.496	0.024	0.164
	H_2O	**0.824**	−0.392	0.076	0.327
	NO_2	0.037	**−0.946**	0.094	0.219
	Eigen value	2.822	1.317	0.992	0.68
	% of variance	42.447	27.332	19.842	7.595
	Cumulative %	42.447	69.779	89.621	97.216
SWM season	CO_2	−0.352	−0.054	0.024	0.1
	CH_4	0.25	0.444	0.173	0.105
	H_2O	**0.931**	−0.175	0.102	0.112
	NO_2	0.152	**−0.849**	0.248	0.124
	Eigen value	2.925	1.188	0.96	0.76
	% of variance	43.495	24.751	19.208	9.339
	Cumulative %	43.495	68.246	87.454	96.793

transformed columnar ozone were selected. Table 3 summarize the results of the data analysis for the NEM and SWM seasons, respectively.

4.3 Model fitting

The previous section of the paper discussed the selection of a subset of the predictor variables that resulted in the best-fitted columnar ozone regression equation through MRA. The criteria used to select the original independent variables included high loading for each PC to generate a regression equation with a high coefficient of determination after substituting the values (Al-Alawi et al., 2008). For both the NEM and SWM seasons, Tables 2 and 3 were used to match a PC in the regression analysis with independent variables. [H_2O] and [NO_2] were selected from PC1 and PC2, respectively, and these two variables function as predictor variables in the subsequent regression analysis. The following columnar ozone models of the NEM and SWM seasons were derived (in Dobson units).

For the NEM season:

$$(\mathrm{PCA1})\mathrm{O_3} = 251.152 - 0.928\mathrm{H_2O} + 0.102\mathrm{NO_2}. \quad (1)$$

For the SWM season:

$$(\mathrm{PCA2})\mathrm{O_3} = 258.822 - 0.872\mathrm{H_2O} + 0.218\mathrm{NO_2}. \quad (2)$$

4.4 Comparison and validation of regression equations

Validation was conducted between the predicted columnar ozone regression equations and the observed columnar ozone values obtained from the SCIAMACHY. To evaluate the columnar values of ozone from Eqs. (1) and (2), we conducted a linear regression correlation of two months selected from (PCA1) for NEM and from (PCA2) for SWM in 2009 for all of Peninsular Malaysia. Validation was also conducted with a linear regression correlation for all the months of 2009 over Petaling Jaya Station. Figure 2 shows the two months selected from the NEM and SWM seasons in 2009, respectively: 1) March and April and 2) July and August. The results showed that the predicted columnar ozone regression equation yielded a strong correlation coefficient for both the NEM and SWM season models using two variables, [H_2O] and [NO_2]. The adjusted coefficients (R^2) were 0.769, 0.789, 0.7, and 0.802 for March, April, July, and August, respectively. Furthermore, a high adjusted coefficient (R^2) of approximately 0.852 was obtained for Petaling Jaya Station.

The coefficients of the regressions were highly significant, and the ρ-values of these coefficients were < 0.05 ($\rho < 0.05$). SPSS software was used to perform the relevant statistical analysis and generate the ρ-value associated with the test statistic and a confidence interval for the mean difference by an independent samples t-test. The highly correlated results indicated that the predicted regression equations for columnar ozone are accurate, so the predicted ozone from the PCA in this study is quite similar to the observed ozone in 2009 for both the NEM and SWM seasons. The columnar ozone predicted by Eqs. (1) and (2) was plotted against the observed value from the SCIAMACHY (Fig. 2).

The monthly differences between the columnar ozone values observed by the SCIAMACHY, predicted from the model, and measured by the Atmospheric Infrared Sounder (AIRS) and *in situ* at Petaling Jaya Station from January to December 2009 are shown in detail in Fig. 3. Figure 3 also shows that the predicted values can exhibit the same pattern as the observed values; predicted monthly columnar ozone is quite consistent with observed columnar ozone for most of the months in 2009. Slight discrepancies between the predicted and the observed values and the observed means and *in situ* measurements were observed during the NEM season. These findings could be attributed to factors not considered in this study, including meteorological parameters (relative humidity, wind speed, precipitation, and temperature).

The predicted curve also shows ripples that parallel the *in situ* measurements throughout the year. Columnar ozone was predicted to determine the accuracy of our proposed model, and the result revealed a close compatibility between the predicted columnar ozone values and the observed mean from the *in situ* measurements at Petaling Jaya Station. This finding shows the accuracy and efficiency of both predicted PCA1 and PCA2 regression equations generated using the combination of MRA and PCA methods to predict columnar ozone in Peninsular Malaysia.

5 Conclusions

Ozone is normally present in the atmosphere in very low quantities, but as it becomes more prevalent, it is

Table 3 Linear regression model used to predict columnar ozone using principal components

Season	Predictors	Constant	PC1	PC2
NEM season	Adjusted R^2		0.859	0.867
	Estimated regression coefficient	251.152	0.928	0.718
SWM season	Adjusted R^2		0.757	0.802
	Estimated regression coefficient	258.822	0.872	0.691

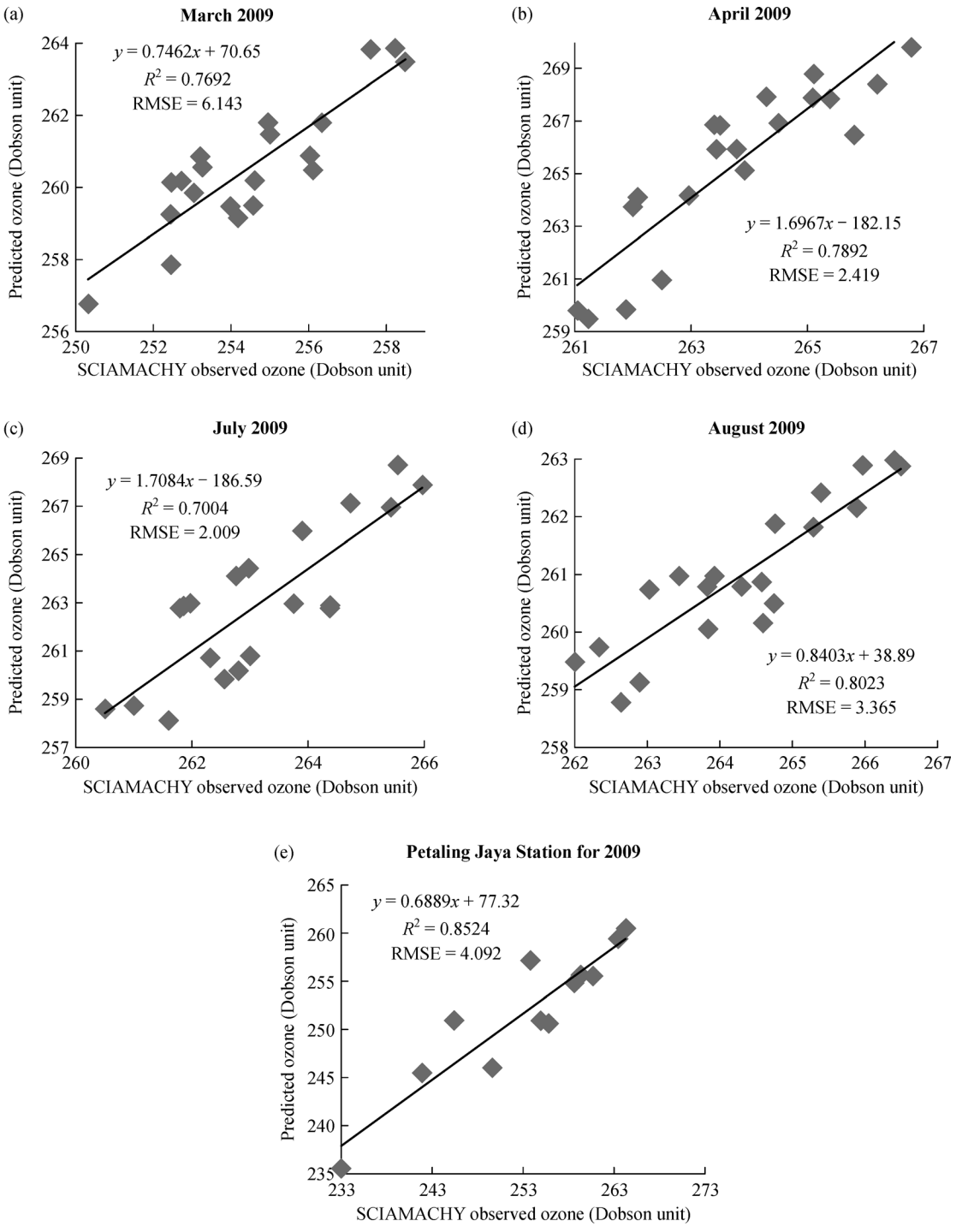

Fig. 2 Predicted versus observed columnar ozone values for (a) March, (b) April, (c) July, (d) August, and (e) Petaling Jaya Station for 2009.

considered to be one of the main pollutants affecting human health and the environment. This study provides information regarding the total columnar ozone in Peninsular Malaysia from 2003 to 2008 and provides the basis of a short-term forecasting tool, which uses other atmospheric variables as predictors that can be used to accurately predict columnar ozone. Six years of satellite data (2003 to 2008) were used to develop the regression equations and calculate the columnar ozone over Peninsular Malaysia during the NEM and SWM seasons using both PCA and MRA methods.

Columnar ozone was negatively correlated with H_2O but positively correlated with CO_2 and NO_2 during the NEM and SWM seasons from 2003 to 2008. This result was

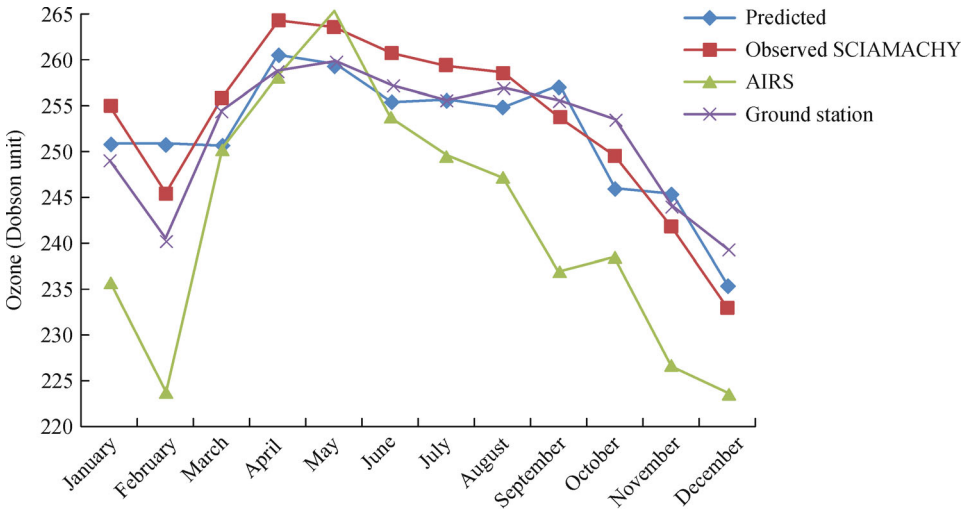

Fig. 3 Comparison of the observed ozone from the SCIAMACHY, the predicted ozone, the AIRS measurements, and the *in situ* measurements at Petaling Jaya Station in 2009.

expected because NO_2 is a precursor of ozone, so an increase in columnar ozone concentration was associated with an increase in NO_2 and a decrease in the level of H_2O. In the NEM season, the columnar ozone was negatively correlated with H_2O (−0.847), NO_2 (0.754), and CO_2 (0.477); it was also negatively but weakly correlated with CH_4 (−0.035). In the SWM season, columnar ozone was highly positively correlated with NO_2 (0.855), CO_2 (0.572), and CH_4 (0.321) and highly negatively correlated with H_2O (−0.832). The regression equations that best fit the columnar ozone data were determined using four independent variables. Our results yielded similar R ($\approx$ 0.83) values for both the NEM and SWM seasons.

The columnar ozone values from the SCIAMACHY were compared with the *in situ* data from Petaling Jaya in 2009 to evaluate the accuracy of the predictive columnar ozone model. The predicted monthly columnar ozone values are quite consistent with the observed columnar ozone data from the SCIAMACHY as well as with the *in situ* data for most of the months in 2009. Validation between the predicted columnar ozone and the observed ozone from the SCIAMACHY was conducted through linear regression correlation. The validation resulted in high correlation coefficients (R = 0.837 to 0.923) and adjusted coefficients (R^2 = 0.700 to 0.852), indicating that the model is accurate and efficient. The validations and comparisons successfully demonstrate the high accuracy of the regression equation.

In summary, this study is the first to use the SCIAMACHY data to analyze the impacts of atmospheric variables on columnar ozone, and it resulted in an accurate prediction model of columnar ozone by combining MRA and PCA methods. Thus, the results clearly indicate the benefit of using satellite SCIAMACHY data to identify the effects of atmospheric variables on columnar ozone over Peninsular Malaysia. The comparison and validation performed in this study support the high accuracy of the regression equations for both the NEM and SWM seasons.

The present study involved the short-term prediction of O_3 modeling in Peninsular Malaysia, which is the western part of Malaysia. Thus, there is a need to further consider the implications of varying data quality. For example, the reanalysis data represents the best currently available observational data, which provides global coverage with more than twenty years of temporal coverage. In addition, the same statistical methods can be applied in the eastern part of Malaysia to analyze atmospheric variables and generate a new algorithm to predict O_3 in the future. In addition, the study can be extended by applying the same study methods in different study areas, such as Southeast Asia.

Acknowledgements The authors would like to thank the Institute of Environmental Physics at the University of Bremen for providing the SCIAMACHY data. This project was carried out with financial support from the Investigation of the Impacts of Summertime Monsoon Circulation to the Aerosols Transportation and Distribution in Southeast Asia which can Lead to Global Climate Change RUI, 1001/PFIZIK/811228, and the Environmental Effects and its Influence on Increased Greenhouse Gases in Peninsular Malaysia Science Fund, 305/PFIZIK/613615. The authors would also like to extend their gratitude to USM for providing support and encouragement and to the USM technical staff for their unwavering support and cooperation.

References

Abdul-Wahab S A, Bakheit C S, Al-Alawi S M (2005). Principle component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ Model Softw, 20(10): 1263–1271

Ahammed Y N, Reddy R R, Gopal K R, Narasimhulu K, Basha D B, Reddy L S S, Rao T V R (2006). Seasonal variation of the surface ozone and its precursor gases during 2001–2003, measured at

Anantapur (14.628N), a semi-arid site in India. Atmos Res, 80(2–3): 151–164

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, 20: 1263–1271

Azid A, Juahir H, Toriman M E, Kamarudin M K A, Saudi A S M, Hasnam C N C, Aziz N A A, Azaman F, Latif M T, Zainuddin S F M, Osman M R, Yamin M (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. Water Air Soil Pollut, 225(8): 2063

Baker D J, Richards G, Grainger A, Gonzalez P, Brown S, Defries R, Held A, Kellndorfer J, Ndunda P, Ojima D, Skrovseth P E, Souza C Jr, Stolle F (2010). Achieving forest carbon information with higher certainty: a five-part plan. Environ Sci Policy, 13(3): 249–260

Bian J, Gettelman A, Chen H, Pan L (2007). Validation of satellite ozone profile retrievals using Beijing ozonesonde data. J Geophys Res, 112 (D6): D06305

Bracher A, Lamsal L N, Weber M, Bramstedt K, Coldewey-Egbers M, Burrows J P (2005). Global satellite validation of SCIAMACHY O_3 columns with GOME WFMDOAS. Atmos Chem Phys, 5(9): 2357–2368

Buchwitz M, de Beek R, Burrows J P, Bovensmann H, Warneke T, Notholt J, Meirink J F, Goede A P H, Bergamaschi P, Körner S, Heimann M, Schulz A (2005). Atmospheric methane and carbon dioxide from SCIAMACHY satellite data: initial comparison with chemistry and transport models. Atmos Chem Phys, 5(4): 941–962

Coldewey-Egbers M, Weber M, Lamsal L N, de Beek R, Buchwitz M, Burrows J P (2005). Total ozone retrieval from GOME UV spectral data using the weighting function DOAS approach. Atmos Chem Phys, 5(4): 1015–1025

Dueñas C, Fernández M C, Cañete S, Carretero J, Liger E (2004). Analyses of ozone in urban and rural sites in Málaga (Spain). Chemosphere, 56(6): 631–639

Johnson C E, Stevenson D S, Collins W J, Derwent R G (2002). Interannual variability in methane growth rate simulated with a coupled Ocean-Atmosphere-Chemistry model. Geophys Res Lett, 29 (19): doi: 10.1029/2002GL015269

Lengyel A, Héberger K, Paksy L, Bánhidi O, Rajkó R (2004). Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere, 57(8): 889–896

Lin W, Xu X, Zhang X, Tang J (2008). Contributions of pollutants from North China Plain to surface ozone at the Shangdianzi GAW Station. Atmos Chem Phys, 8(19): 5889–5898

Mieruch S, Noël S, Bovensmann H, Burrows J P (2008). Analysis of global water vapour trends from satellite measurements in the visible spectral range. Atmos Chem Phys, 8(3): 491–504

Omar D (2009). Urban form and sustainability of a hot humid city of Kuala Lumpur. J Soc Sci, 8: 353–359

Pochanart P, Kreasuwun J, Sukasem P, Geeratithadaniyom W, Tabucanon M S, Hirokawa J, Kajii Y, Akimoto H (2001). Tropical tropospheric ozone observed in Thailand. Atmos Environ, 35(15): 2657–2668

Rajab J M, MatJafri M Z, Lim H S (2013). Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia. Atmos Environ, 71: 36–43

Reddy B S K, Kumar K R, Balakrishnaiah G, Gopal K R, Reddy R R, Sivakumar V, Lingaswamy A P, Arafath S Md, Umadevi K, Kumari S P, Ahammed Y N, Lal S (2012). Analysis of diurnal and seasonal behaviour of surface ozone and its precursor (NO_x) at a semi-arid rural site in Southern India. Aerosol Air Qual Res, 12: 1081–1094

Richter A, Burrows J P, Nüß H, Granier C, Niemeier U (2005). Increase in tropospheric nitrogen dioxide over China observed from space. Nature, 437(7055): 129–132

Richter A, Eyring V, Burrows J P, Bovensmann H, Lauer A, Sierk B, Crutzen P J (2004). Satellite measurements of NO_2 from international shipping emissions. Geophys Res Lett, 31(23): L23110

Schneising O, Buchwitz M, Burrows J P, Bovensmann H, Bergamaschi P, Peters W (2008a). Three years of greenhouse gas column averaged dry air mole fractions retrieved from satellite-part 2: methane. Atmos Chem Phys, 8(3): 8273–8326

Schneising O, Buchwitz M, Burrows J P, Bovensmann H, Reuter M, Notholt J, Macatangay R, Warneke T (2008b). Three years of greenhouse gas column-averaged dry air mole fractions retrieved from satellite – Part 1: carbon dioxide. Atmos Chem Phys, 8(14): 3827–3853

Statheropoulos M, Vassiliadis N, Pappa A (1998). Principle component and canocical correlation analysis for examining air pollution and meteorological data. Atmos Environ, 32(6): 1087–1095

Stephens B B, Gurney K R, Tans P P, Sweeney C, Peters W, Bruhwiler L, Ciais P, Ramonet M, Bousquet P, Nakazawa T, Aoki S, Machida T, Inoue G, Vinnichenko N, Lloyd J, Jordan A, Heimann M, Shibistova O, Langenfelds R L, Steele L P, Francey R J, Denning A S (2007). Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO_2. Science, 316(5832): 1732–1735

Tan K C, Lim H S, Mat Jafri M Z (2012a). Total ozone column distribution over Peninsular Malaysia from Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY). Proceeding SPIE 8538, Earth Resources and Environmental Remote Sensing/GIS Applications III, 85380Y

Tan K C, Lim H S, Mat Jafri M Z (2012b). Satellite observation distribution of atmospheric ozone over Peninsular Malaysia from SCIAMACHY. International Coference on Control System, Computing and Engineering (ICCSCE 2012), 238–243

Tan K C, Lim H S, Mat Jafri M Z (2013). Relationship between ozone and their air pollutants in Peninsular Malaysia for 2003 retrieved from SCIAMACHY. AIP Conf Proc, 1528: 151–156

Tan K C, Lim H S, Mat Jafri M Z (2014a). Analysis of total column ozone in Peninsular Malaysia retrieved from SCIAMACHY. Atmos Pollut Res, 5(1): 42–51

Tan K C, Lim H S, Mat Jafri M Z (2014b). Multiple regression analysis in modeling of columnar ozone in Peninsular Malaysia. Environ Sci Pollut Res Int, 21(12): 7567–7577

Tan K C, Lim H S, Mat Jafri M Z. (2012c). Total ozone column distribution over Peninsular Malaysia from scanning imaging absorption spectrometer for atmospheric cartography (SCIAMACHY). Proc. SPIE 8538, Earth Resources and Environmental Remote Sensing/GIS Applications III, 85380Y (October 25, 2012),

Tian W, Chipperfield M P, Lü D (2009). Impact of increasing stratospheric water vapor on ozone depletion and temperature change. Adv Atmos Sci, 26(3): 423–437

Toh Y Y, Lim S F, Von Glasow R(2013). The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah Rata, Malaysia. Atmos Environ, 70: 435–446

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

Vingarzan R (2004). A review of surface ozone background levels and trends. Atmos Environ, 38(21): 3431–3442

Wang Y S, Zhou L, Wang M X, Zheng X H (2001). Trends of atmospheric methane in Beijing. Chemosphere, Glob Chang Sci, 3 (1): 65–71

Wu H W Y, Chan L Y (2001). Surface ozone trends in Hong Kong in 1985–1995. Environ Int, 26(4): 213–222

Yonemura S, Tsuruta H, Kawashima S, Sudo S, Leong C P, Lim S F, Zubaidi J, Masayasu H (2002). Tropospheric ozone climatology over Peninsular Malaysia from 1992 to 1999. J Geophys Res, 107(D15): 4229

AUTHOR BIOGRAPHIES

Tan Kok Chooi is a Senior Lecturer at the School of Physics, Universiti Sains Malaysia. He obtained his Ph.D in image processing from Universiti Sains Malaysia. His research interests cover the field of remote sensing applications and environmental monitoring, such as land use/land cover changes, land surface properties, and image classification. Currently, his research focuses on the environmental remote sensing, atmospheric chemistry and physics. E-mail: kctan@usm.my.

Lim Hwee San is an Associate Professor of Geophysics and Remote Sensing at the School of Physics, Universiti Sains Malaysia. He obtained his Ph.D from Universiti Sains Malaysia in environmental remote sensing. His research interests lie generally in the field of optical remote sensing- both passive and active-and digital image processing, particularly as it applies to spectral image data. In both cases, the primary applications are water quality monitoring, air quality monitoring, land cover/change detection, land surface properties and digital images classification. He also interested in modelling of the optical properties of atmospheric aerosols. His current effort focus on the applications of ground based LIDAR and satellite based LIDAR (e.g., CALIPSO, AIRS) data for air pollution and green house effects study. E-mail: hslim@usm.my.

Mohd. Zubir Mat Jafri is a Professor of Optical Sensing and Remote Sensing at the School of Physics, Universiti Sains Malaysia. He obtained his Ph.D from Universiti College of Swansea, Wales, UK, in digital image processing. He has more than twenty years teaching and research experience in the area of physics, optical communication, digital image processing, electronics and microprocessor technology. E-mail: mjafri@usm.my.