Atmospheric Pollution Research 7 (2016) 533–546

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Atmospheric Pollution Research

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

Original article

Prediction of column ozone concentrations using multiple regression analysis and principal component analysis techniques: A case study in peninsular Malaysia

Kok Chooi Tan*, Hwee San Lim, Mohd Zubir Mat Jafri

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

A R T I C L E I N F O

Article history:
Received 15 September 2015
Accepted 9 January 2016
Available online 2 February 2016

Keywords:
Ozone
Multiple regression analysis
Principal component analysis
Peninsular Malaysia
Correlation coefficient

A B S T R A C T

The aim of this study is to develop new algorithms of the column ozone (O_3) in Peninsular Malaysia using statistical methods. Four regression equations, denoted as O_3 NEM, O_3 SWM, (PCA1) O_3 NEM season, and (PCA2) O_3 SWM season, were developed. Multiple regression analysis (MRA) and principal component analysis (PCA) methods were utilized to achieve the objectives of the study. MRA was used to generate regression equations for O_3 NEM and O_3 SWM, whereas a combination of the MRA and PCA methods were used to generate regression equations for PCA1 and PCA2. The results of the best regression equations for the column O_3 through MRA by using four of the independent variables were highly correlated ($R = 0.811$ for SWM, $R = 0.803$ for NEM) for the six-year (2003–2008) data. However, the result of fitting the best equations for the O_3 data using four of the independent variables gave approximately the same R values (≈ 0.83) for both the NEM and SWM seasons using the combined MRA and PCA methods. The common variables that appeared in both regression equations were H_2O vapor and NO_2. This result was expected because NO_2 is a precursor of O_3. The correlation coefficients (R) of the validation for the NEM and SWM seasons were 0.877–0.888 and 0.837–0.896, respectively. These statistical values indicated a very good agreement between the monthly predicted and observed O_3 for Peninsular Malaysia.

1. Introduction

Ozone (O_3) is secondary pollutant formed by a photochemical reaction among volatile organic compounds (VOCs) and nitrogen oxides (NO_x) in the atmosphere (Castell-Balaguer et al., 2012; Tan et al., 2014a). O_3 plays an important role in the atmospheric process and influences the environment and human health depending on its location in the atmosphere. O_3 in the troposphere is the third most important greenhouse gas and plays a crucial role in atmospheric chemistry by controlling the oxidizing capacity of the atmosphere (Nishanth et al., 2012). Inversely, O_3 in the stratosphere filters out harmful ultra-violet radiation from the sun. In addition, O_3 is an important heating source in the stratosphere (Park et al., 2012).

O_3 is a gas present in the atmosphere in a very small quantity compared with other gases. However, O_3 is an important chemical constituent of the atmosphere, playing a key role in the atmospheric energy budget and chemistry, air quality and global change (Dueñas et al., 2004; Ahammed et al., 2006; Lin et al., 2008). Moreover, O_3 also imparts significant impacts on the radiation budget of the atmosphere (Wu and Chan, 2001). Hence, as a pollutant and greenhouse gas, abrupt changes in atmospheric O_3 due to high levels of surface O_3 and anthropogenic emissions may create environmental problems and contribute to climate change (Guicherit and Roemer, 2000; Vingarzan, 2004). In addition, the global issue in the field of climate and environmental science has become important since the discovery of the O_3 hole in Antarctica (Farman et al., 1985). The research primarily focuses on the recovery and stratospheric O_3 trend, thus prompting more examination into the spatial and temporal variability of O_3 at different locations around the globe (Oluleye and Okogbue, 2013).

* Corresponding author. Tel.: +60 4 6535326; fax: +60 4 6579150.
E-mail addresses: kctan@usm.my (K.C. Tan), hslim@usm.my (H. San Lim), mjafri@usm.my (M.Z.M. Jafri).

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control (TUNCAP).

http://dx.doi.org/10.1016/j.apr.2016.01.002

Besides, O_3 also formed in the atmosphere as the result of the burning of biomass as biomass burning leads to the formation of aerosol particles and the combustion products are O_3 precursors (Toh et al., 2013). Previously, few studies showed that the elevated O_3 concentrations were due to biomass emissions from regions far from the origin of the sources (Boian and Kirchhoff, 2005; Shan et al., 2009). Thus, Southeast Asia haze and pollution contributes to local air pollution and extends to a regional and global scale. For example, the biomass burning in Indonesia is attributed with causing the tremendous amount of dense and thick smoke that affects neighboring countries such as Malaysia. Kita et al. (2000) suggested that the total O_3 column over the Indonesian region increased repeatedly during 1982–1983, 1987, 1991, 1994, and 1997, caused by the active forest fires during the dry season of the El Niño periods. Page et al. (2002) estimated that the forest fires in Indonesia during 1997 generated the equivalent to 13–40% of the mean annual global carbon emissions from fossil fuel. Hyer and Chew (2010) utilized the Navy Aerosol and Analysis System (NAAPS) and Fire Locating and Monitoring of Burning Emissions (FLAMBE) smoke flux model to observe the extreme pollution events in Singapore and Malaysia during September–November 2006. The results indicated that burning biomass smoke from fires in Sumatra and Borneo contributed to almost all of the extreme PM_{10} periods during the observation time over the entire region. Sahani et al. (2014) examined the risk of haze days due to biomass burning in Southeast Asia on the daily mortality in the Klang Valley region between 2000 and 2007. The results indicated that each day of haze in the Klang Valley accounted for an immediate effect of respiratory mortality with a 19% increase in respiratory mortality, and 34% and 41% increase in respiratory mortality among all males and elderly male residents, respectively. Additionally, the delayed effects of haze were found to affect the respiratory mortality for adult females aged 15–59 years old and the natural mortality of children, with increases of 66% and 41.4%, respectively.

The effects of O_3 were revealed by Middleton more than 50 years ago (Lehman et al., 2004). On a scale of 100 years, the increasing trends of tropospheric O_3 are qualitatively in agreement with the emission trends of precursors. This phenomenon has led to the increase in the tropospheric O_3 level becoming one of the most crucial environmental problems requiring a resolution in the coming decades due to its adverse effects (Guicherit and Roemer, 2000; Shan et al., 2008). Traditionally, air quality monitoring stations provided growth-truth data for the monitoring of atmospheric gases, especially in industrial, urban and suburban areas (Hyer and Chew, 2010). Thus, these monitoring stations are located sparsely and it is impossible to acquire data through these stations (Khattak et al., 2014). In the more recent decades, the availability of satellite instruments has allowed atmospheric O_3 monitoring at the global scale to become possible. For example, satellite instruments provide continuous monitoring of various trace and greenhouse gases globally, such as the Global Ozone Monitoring Experiment (GOME, Thomas et al., 2005), Ozone Monitoring Instrument (OMI, Krotkov et al., 2006), Global Ozone Monitoring Experiment-2 (GOME-2), and Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY, Lee et al., 2008).

The analysis and forecasting of air quality parameters are an important issue that requires immediate and serious attention, especially in atmospheric and environmental research, because it is one of the most important factors that affect the quality of living and life (Mahapatra, 2010; Azid et al., 2014). Thus, time series analysis has been widely used in prediction and forecasting studies. Typically, we test and predict the known observations from the past by using them as inputs in a model, to figure out how well the output matches the known observations. Various statistical analyses of long-term total column O_3 records have been performed to examine the effect of external variables on the total O_3 using ground-based measurements (Wohltmann et al., 2007) and/or satellite measurements (Brunner et al., 2006). Ground-based measurements have an advantage as they often span time periods longer than those available from satellite measurements (Brunner et al., 2006). However, satellite instruments perform measurements at a higher temporal frequency (daily) and provide global coverage (Knibbe et al., 2014). Statistical methods have been developed to explain and predict the O_3 concentrations (Tsakiri and Zurbenko, 2011). A wide range of methods have been used to predict the presence of air pollutants (Ghazali et al., 2010).

Among the most frequently used methods are neural network models, time series, and regression. Multiple regression analysis (MRA) is one of the most frequently used methodologies to determine the dependence of a response variable on several independent variables. MRA is usually applied to search for the dependence of a response variable on several independent variables and 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 such multicollinearity; in particular, multivariate data analysis (MDA) is an effective alternative technique to address this drawback. MDA techniques have been widely used for 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 commonly used techniques in air quality studies to analyze voluminous environmental data (Vaidya et al., 2000). In previous studies, PCA methods were successfully used to identify important factors influencing ozone concentrations and examine ozone variations (Lengyel et al., 2004).

Even though an increasing number of studies on surface O_3 in Malaysia have been reported, limited studies have been conducted on the total column O_3 over the vast landmass of Peninsular Malaysia. Malaysia has experienced economic growth due to urbanization, industrialization, and rapid traffic growth, which has resulted in a drastic increase in the emissions of air pollutants into the atmosphere (Streets and Waldhoff, 2000). Heavy pollution is being created by emissions that primarily result from increases in the volume of motor vehicles, industrial areas, and trans-boundary pollution. The high resolution that is associated with special satellite specifications is required to study how atmospheric parameters (including greenhouse gases) affect atmospheric O_3. Due to the lack of observational studies of greenhouse gases in Malaysia, most studies have depended on ground station data. Malaysia, a tropical country in Southeast Asia, has undergone rapid economic development and urban expansion, and transportation facilities have led to increased fossil fuel consumption over the past few decades, which have resulted in the increased emission of air pollutants, especially in industrial areas and cities. Therefore, there is a need to focus the study on a regression analysis between selected atmospheric parameters and column O_3.

2. Data and methodology

2.1. Site description

Peninsular Malaysia is located at latitudes 1°–7° north and longitudes 99°–105° east 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).

Because of its geographical location near the equator, Peninsular Malaysia experiences a humid tropical climate throughout the year;