View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Atmospheric Pollution Research 6 (2015) 245-253

Atmospheric Pollution Research

www.atmospolres.com

# Prediction of surface ozone episodes using clusters based generalized linear mixed effects models in Houston–Galveston–Brazoria area, Texas

**Wei Sun [1], Ahmet Palazoglu [2], Angadh Singh [2], Hao Zhang [3], Qi Wang [1], Zemeng Zhao [1], Ding Cao [4]**

*[1] Beijing Key Lab of Membrane Science and Technology, College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China*
*[2] Department of Chemical Engineering and Materials Science, University of California at Davis, USA*
*[3] School of Chemistry and Chemical Engineering, Southwest University, 2 Tiansheng Road, BeiBei District, Chongqing 400715, China*
*[4] State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, China*

## ABSTRACT

A two–stage strategy is proposed to predict regional peak ozone episodes in the Houston–Galveston–Brazoria (HGB) area of Texas, USA. With the forecasted meteorological information, ozone episodes can be predicted one day in advance. Three generalized linear mixed effects models (GLMMs) are built with air quality and meteorological data monitored at CAMS35, CAMS403 and CAMS1015; wind field data from 8 monitoring sites in HGB area are used to generate clusters which represent distinct weather patterns. Air quality and meteorological data during ozone seasons (Apr. 1st – Oct. 31st) from 2003 to 2005 are used to build site–specific prediction models. Data of ozone season from 2006 to 2007 are used to test these models. Compared to linear regression models (LM), generalized linear models (GLMs), multilayer perceptron (MLP) and support vector machine (SVM), GLMM which considers differences in ozone formation and diffusion in distinct weather patterns has the smallest fitting and prediction error on ozone exceedances and can detect the most number of exceedance days correctly.

***Keywords:*** *Cluster Analysis, generalized linear mixed effects models, random effects, ozone exceedances*

***Corresponding Author:***
*Hao Zhang*
☎ : *+86-185-8048-1938*
📠 : *+86-010-6442-3811*
✉ : *haozhang@swu.edu.cn*

***Article History:***
*Received: 09 June 2014*
*Revised: 17 September 2014*
*Accepted: 20 September 2014*

**doi:** 10.5094/APR.2015.029

## 1. Introduction

As one of the most harmful secondary air pollutants, surface ozone is mainly formed by the photochemical reactions between nitrogen oxides ($NO_X$) and volatile organic compounds (VOCs) (Jenkin and Clemitshaw, 2000). Under certain weather conditions, surface ozone can accumulate to unhealthy levels and cause chronic damages on human respiratory systems (Ji et al., 2011; Groves et al., 2012). Exposure to prolonged high levels of ozone can lead to more visits to physicians and emergency rooms (Fauroux et al., 2000; Tolbert et al., 2000). To alleviate the negative effects of ozone action days (OADs), many prediction strategies have been carried out, especially in the Houston–Galveston–Brazoria (HGB) area of Texas, USA. The benefits to human health by correctly forecasting OADs and issuing health warnings can be significant (Neidell, 2010). Here, days in which the maximum 8 h–average ozone concentration is over 75 ppb are defined as ozone action days (OADs).

Forecasting surface ozone concentration is challenging due to the complexity of the physical and chemical processes involved in photochemical reactions. Various techniques, which could be basically classified as deterministic models and statistical models, have been developed for surface ozone predictions. Deterministic models, which are also named as chemical transport models (CTMs), comprise numerical models which typically simulate atmospheric chemistry and dispersion models (Jacob, 1999). The most popular CTMs include CMAQ, WRF/Chem, CAMx, GEOS–CHEM, MOZART, and others (Done et al., 2004; Tesche et al., 2006; Henze et al., 2007; Emmons et al., 2010). During the Texas Air Quality Study campaigns in 2000 and 2006, several CTMs were used to simulate summer ozone episodes over the HGB area (Jiang and Fast, 2004; Fast et al., 2006; Byun et al., 2007). Simulation results showed that CTMs were sensitive to initial meteorological factors, horizontal grid spacing, nesting methods and land–surface models used in the model simulations due to the complex mathematical structure (Zhang et al., 2007; Misenis and Zhang, 2010). Therefore, CTM parameters may need to be reset frequently and computations may require vast resources. Moreover, the accuracy of OADs predicted by CTMs is relatively low. A rigorous comparison between CTMs and statistical models showed that the correlation between measured daily maximum and surface ozone concentrations yielded using complex 3–D modeling approach is 0.49, which is much lower than that of the site–specific, well–developed data–driven models (Draxler, 2000). The true positive rate (TPR) of ozone action days during 2013 ozone season in Houston announced by Texas Commission on Environment Quality (TCEQ) is only 64%. Compared to CTMs, statistical models based on historical data provide more accurate results and involve simpler calculations. Statistical models used for ozone prediction in HGB area include linear regression, generalized additive models (GAMs), Box–Jenkins ARIMA, clustering and artificial neural networks (ANNs) (Davis and Speckman, 1999; Prybutok et al., 2000; Darby, 2005). Other statistical models such as support vector machines, fuzzy inference systems, evolutionary algorithms and ensembles of predictors also are used to forecast surface ozone levels frequently (Sfetsos and

*© Author(s) 2015. This work is distributed under the Creative Commons Attribution 3.0 License.*

Siriopoulos, 2004; Feng et al., 2011; Sfetsos et al., 2013; Zahedi et al., 2014; Zhang et al., 2014). Among these statistical models, ANNs could predict daily maximum surface ozone levels with the smallest errors. However, ANNs are developed with a non–explanatory structure and are black box approaches and the successful implementation of ANN–based models depend on the proper selection of training data, network structure and connection mode among neural nodes (Psichogios and Ungar, 1992). Though well–developed statistical models outperform CTMs to some extent, they substantially underestimate surface ozone concentrations in OADs which are the most harmful from human health perspective (Cobourn and Hubbard, 1999; Draxler, 2000).

It is essential to develop new prediction models to improve true positive rate (TPR) of OADs. To address this problem, non–Gaussian distributed characteristics of model data should be considered first. Meteorological and air quality data usually are sparse and limited in collected data set (Zhang and Fan, 2008). Ozone action days which take a small fraction of the total modeling data reside at the right end of the probability density curve. Also, some key meteorological factors which affect the photochemical reactions distribute non–normally, such as wind speed, solar radiation and surface temperature (Sun et al., 2013). Thus, a linear regression model might not capture the underlying complex features. In order to reflect correlations between non–normal distributed ozone and independent variables, a GAM has been applied to predict maximum and 8–hr average ozone in Houston but did not attain the expected predictive capabilities (Davis and Speckman, 1999). Thus, only non–Gaussian distribution of model data may not be enough to map complex relations between surface ozone and its influence factors. Differences of surface ozone formation and transport in distinct weather patterns may also be considered in the prediction models. Davis et al. (1998) used clustering to identify seven distinct meteorological regimes and build GAMs in each regime to model meteorological effects on ozone formation. Their results show that meteorological effects on ozone vary significantly in different weather patterns. Darby (2005) also used clustering techniques to study surface winds effects on ozone formation during TexAQS 2000 and found high ozone was most likely to occur with clusters representing the Gulf breeze. Thus, data of ozone seasons can be treated as grouped data and relations between surface ozone and meteorological factors show marked differences. Due to the limited size of model data, generalized linear models based on different subsets of the model dataset with very limited data size may not be reliable. To consider both non–normal distributed and grouped data structure, cluster based GLMM are introduced to predict surface ozone levels in HGB area: with the extension of mixed–effects, GLMMs can be adopted to deal with grouped data which are organized in several clusters; with the link functions, GLMMs can model OADs that reside along the right end of the probability density plot. Therefore, cluster–based GLMMs will be adopted in this work to improve TPR of ozone action days in the HGB area.

In the sequel, first, model data are grouped into several classes according to underlying synoptic wind fields; then, GLMMs are built to predict surface ozone levels 24 hours in advance. The remainder of the paper is organized as follows: The method proposed is described in the second section, which includes a two–step cluster algorithm and GLMM; followed by the introduction of data and the study area. Then result and discussion are addressed in detail. The conclusions are presented in the final section.

## 2. Methodology

### 2.1. Generalized linear mixed model

As GLMM is originated from linear regression, a linear regression model needs to be discussed first. Such a model can be described as follows:

$$y = X\beta + \varepsilon, \varepsilon \sim N_n(0, \sigma^2 I_n) \tag{1}$$

where, $y$ is the response variable, 8–h average surface ozone concentration in this work; $X$ is the input variable ensembles which include ozone precursors and meteorological factors; $\beta$ is the vector of regression coefficients; and $\varepsilon$ is the prediction errors, which follow normal distributions with the mean value of 0 and the variance of $\sigma^2$. In the linear regression model, $X\beta$ can be treated as fixed effects and $\varepsilon$ as the random effect. With fixed effects only, linear regression model is not appropriate to describe modeling data which can be grouped into distinct weather patterns. Thus, random effect terms are used to form a linear mixed model (LMM). With the random effects, LMMs can capture the meteorological variability that drives ozone formation among different weather patterns. A LMM can be described as follows:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, b_i \sim N_q(0, \Psi), \varepsilon_i \sim N_{ni}(0, \delta^2 \Lambda_i) \tag{2}$$

where, $y_i$ is the surface ozone concentration of the $i^{th}$ weather pattern; $X_i$ is the fixed effects for ozone precursors and meteorological factors of the $i^{th}$ weather pattern; $\beta$ is the vector of fixed effect coefficients; $Z_i$ is the random effects, which are used to identify different correlations between surface ozone and its influence factors in the $i^{th}$ weather pattern; $b_i$ is the coefficients of random effect of the $i^{th}$ weather pattern; $\Psi$ is the covariance for the random effects and $\sigma^2\Lambda$ is the covariance for errors in the $i^{th}$ weather pattern.

When the prediction errors distribute non–normally and ozone data follows any distribution belonging to the exponential distribution family, a one–to–one continuous link function $g$ can be used to model the mean value of surface ozone. With the random effects and link function, a GLMM can be described as Bolker et al. (2009):

$$\mu = g^{-1}(\eta_i) = g^{-1}(X_i\beta + Z_i b_i) \tag{3}$$

where, $\mu$ is the mean value of the predicted variable, $\eta$ is the linear combination of $\beta$ and $g$ is the link function.

In this paper, the modeling data are grouped into different clusters according to the characteristics of the diurnal wind cycles. The influence of meteorological factors and ozone precursors on ozone buildup makes a great deal of difference in different synoptic weather patterns. With GLMM, it is possible to account for the influence of different synoptic patterns on ozone dynamics. Moreover, mixed effects modeling approach could make use of the whole data instead of building models with subsets exclusively.

### 2.2. Clustering methods

**Two–stage clustering methods based on k–means and dendrogram.** Usually, high air pollutant concentrations in urban areas do not typically result from emission events but from changes of meteorological conditions. Thus, the clustering algorithm proposed by Beaver and Palazoglu (2006a) is employed to identify wind field patterns which play different roles in ozone formation in HGB area. This method can extract wind field dynamics and recognize surface flow patterns which affect local ozone concentration by grouping days exhibiting similar diurnal cycles. It consists of two steps: the first step is the standard k–means clustering for matrices of wind data time series; the second step is an aggregation step which aggregates cluster solutions generated by performing many randomly initialized runs of the first step into a single, hierarchical solution. Details of this method can be found in the paper by Beaver and Palazoglu (2006b).

**New data labeling.** When this model is used for ozone prediction, wind field of the next day should be labeled at the beginning. First, wind field data are formed into a lagged matrix, whose structure is the same as that used in cluster analysis discussed above. Then the k dynamic principal component analysis (DPCA) prototypes of each run are concatenated into a row:

$$P = [(P_{11}, P_{21} \ldots P_{k1})_{1}, (P_{12}, P_{22} \ldots P_{k2})_{2} \ldots (P_{1n}, P_{2n} \ldots P_{kn})_{n}] \tag{4}$$

Residuals of the new data can be calculated by applying the new data window to the k DPCA prototypes. Then, it is assigned to the prototype which generates the minimum residual and the corresponding element changes into 1. Next a binary matrix which has the same column as $H$ and named $H_{new}$ can be obtained. The new part of distance matrix $D_{new}$ is calculated as:

$$D_{new} = 1 - \frac{1}{n} \left\langle \begin{matrix} [0]_{N\times(k\times n)} \\ H_{new} \end{matrix} \right\rangle \times \left\langle \begin{matrix} [0]_{N\times(k\times n)} \\ H_{new} \end{matrix} \right\rangle^{T} \tag{5}$$

where, [0] is a zero matrix which has the same size as $H$. $D_{new}$ is the new distance matrix which is a $(N+N_{new})\times(N+N_{new})$ square matrix. $N_{new}$ is the number of new windows. From the $N+1^{st}$ row (column) to $N+N_{new}$ row (column), each row (column) describes the dissimilarity of the new window to the old windows. In a standard hierarchical clustering algorithm, new window will be assigned to the cluster in which a certain member has the smallest dissimilarity value to the new window.

## 3. Study Area and Data Characteristics

### 3.1. Study area

The geography and monitoring sites in HGB area are shown in Figure 1. HGB area is a highly populated area and a hub of petroleum extraction and refining industries. According to census data, more than 4 million people reside in HGB area. It has been one of the major air pollution source regions in the United States. A large amount of ozone precursors are emitted by industrial plants, traffic and the ship channel activities (Vizuete et al., 2008). Coupled with warm weather patterns, HGB area suffers from ozone action days frequently. There are two main factors that contribute to HGB's high ozone concentrations. One is the high emission rates of VOCs and $NO_X$ from Houston urban activity, power plants and industrial plants along the Ship Channel, and the other is the land–sea breeze circulation. The condition of boundary layer during daytime is pretty unstable, but remains very stable during night-time (Misenis and Zhang, 2010). According to the results of TexAQS 2000, most ozone action days were associated with the land–sea breeze circulation.

### 3.2. Air quality and meteorological data

The meteorological data used in this paper are downloaded from the website of TCEQ and National Oceanic and Atmospheric Administration (NOAA). Data of ozone season (Apr. 1st – Oct. 31st) from 2003 to 2005 in HGB area are used to build and validate the GLMM. Concentration of air pollutants in the current day which include $O_3$, $NO_X$, CO, organic carbon $PM_{2.5}$ and non–methane hydrocarbons (NMHCs), and meteorological factors announced by NOAA in the next day which include wind speed ($W_x$, $W_y$), maximum wind gust (MWG), outdoor temperature (OT), relative humidity (RH), solar radiation (SR) and ultraviolet radiation ($UR$) are used as predictors of GLMM. To derive general conclusions, three GLMMs are built for monitoring site CAMS35, CAMS403 and CAMS1015. CAMS35 located in a residential area at Deer Park and has been activated since 1996; CAMS403 is adjacent to Clinton Dr. and has been activated since 1972; and CAMS1015 is set beside the Ship Channel and has been activated since 2003. In the wind field clustering step, hourly average wind speed data of 8 stations are used, which are supposed to capture the main wind field features in the HGB area. Details of modeling data and monitoring sites are shown in Tables 1 and 2.

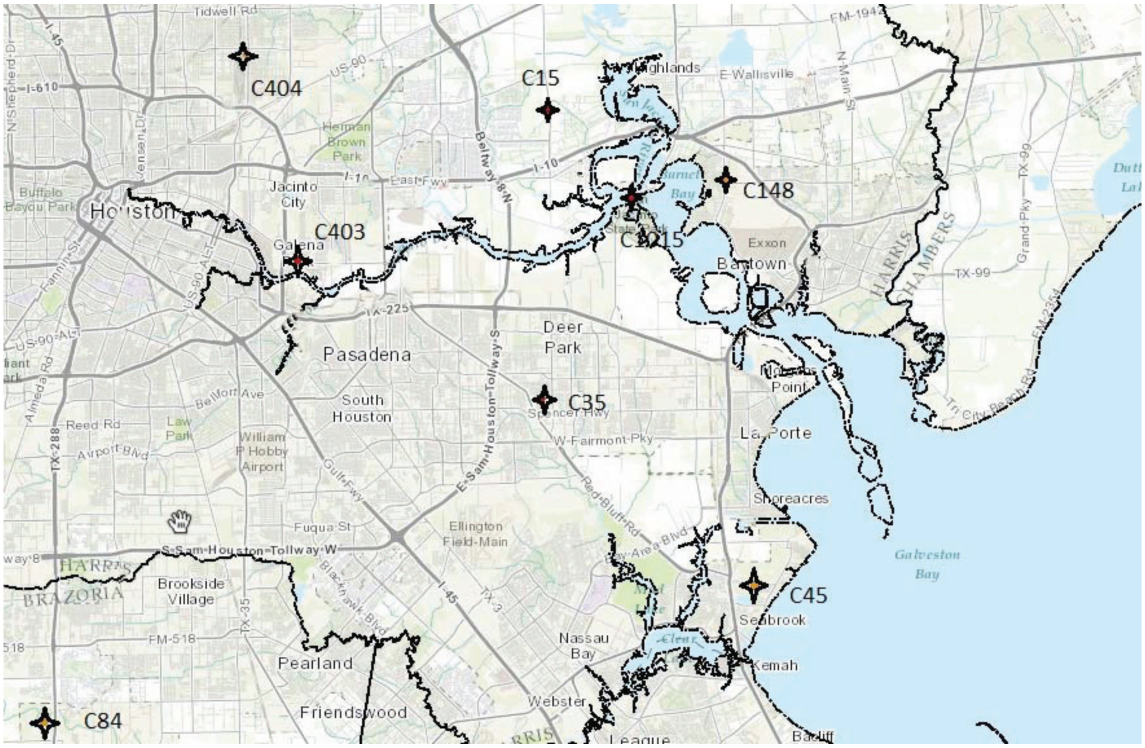

*Figure 1. The geography and monitoring sites in HGB area.*

***Table 1.*** *Details of air pollutants and meteorological factors used in GLMMs for predicting $O_3$ concentrations on the next day*

| Variable Name | Unit | Temporal Frequency | Monitoring Time |
|---|---|---|---|
| $O_3$ | ppb | 8 h average | Current day |
| CO | ppm | 8 h average | Current day |
| NO | ppb | 8 h average | Current day |
| $NO_2$ | ppb | 8 h average | Current day |
| $NO_x$ | ppb | 8 h average | Current day |
| NMHC | ppb | 8 h average | Current day |
| $PM_{2.5}$ OC | $\mu g/m^3$ | 8 h average | Current day |
| *MWG* | mph | 8 h average | Next day |
| $W_x$ | mph | 8 h average | Next day |
| $W_y$ | mph | Hourly average | Next day |
| *OT* | °F | 8 h average | Next day |
| *RH* | | 8 h average | Next day |
| *SR* | Ly/min | 8 h average | Next day |
| *UR* | Ly/min | 8 h average | Next day |

***Table 2.*** *Continuous Air Monitoring Station (CAMS) numbers and site information used in the cluster analysis*

| CAMS No. | Site Description | Address | Activated Year |
|---|---|---|---|
| C15 | Channelview | 1405 Sheldon Road | 1980 |
| C35 | Houston Deer Park 2 | 4514 1/2 Durant St. | 1996 |
| C45 | Seabrook Friendship Park | 4522 Park Rd. | 2001 |
| C84 | Manvel Croix Park | 4503 Croix Pkwy | 2001 |
| C148 | Baytown | 7210 1/2 Bayway Drive | 1998 |
| C403 | Clinton | 9525 1/2 Clinton Dr. | 1972 |
| C404 | Houston Kirkpatrick | 5565 Kirkpatrick | 2000 |
| C1015 | Lynchburg Ferry | 4407 Independence Parkway South | 2003 |

## 4. Results and Discussion

### 4.1. Cluster results for ozone seasons in HGB area

In this work, wind field data during extended ozone season from 2003 to 2005 are used to identify distinct weather patterns which have different effects on ozone formation and dispersion in HGB area. With new data labeling methods, days during the ozone season from 2006 to 2007 are assigned into each cluster determined by data from 2003 to 2005. During the clustering step, wind fields in HGB area are separated into 5 groups, which include two anticyclonic clusters (1 and 2) and three cyclonic clusters (3, 4 and 5). Ozone action days occur frequently in one of the anticyclonic clusters. There are 108 ozone action days recorded by CAMS35 from 2003 to 2007, 53 by CAMS403 and 76 by CAMS1015, respectively. Three ozone action days at CAMS35 are not part of the ozone season (Mar. 23rd, 2003 / Mar. 31st, 2004 / Nov. 4th 2007), one at CAMS403 and one at CAMS1015 (Nov. 6th, 2004). One of the anticyclonic clusters captures most ozone action days. Details of the cluster results are shown in Table 3. The differences of ozone concentration, wind speed in *x* direction, wind speed in *y* direction during nighttime and wind speed in *y* direction during day time in distinct wind patterns are shown in Figure 2. It is shown in Figure 2 that there is the highest surface ozone concentration in cluster 1. Wind speed in *x* direction is the lowest which is favorable for surface ozone accumulation in the first weather pattern. The most notable characteristics of cluster 1 are the reverse of wind direction from nighttime to day time. During night, low speed wind blows from land to the sea and keeps $NO_X$ emitted during nighttime up to a high level. In day time, wind which blows from the sea to the downtown of Houston will carry air rich in precursors of surface ozone to the land.

As cluster analysis cannot deal with data with missing values, an EM imputation method proposed by Schneider (2001) is used to impute modeling data. Days which lack data for more than 8 straight hours are deleted from the modeling dataset. After data imputation, there are 923 days, 816 days and 879 days available for monitoring site CAMS35, CAMS403 and CAMS1015, respectively.

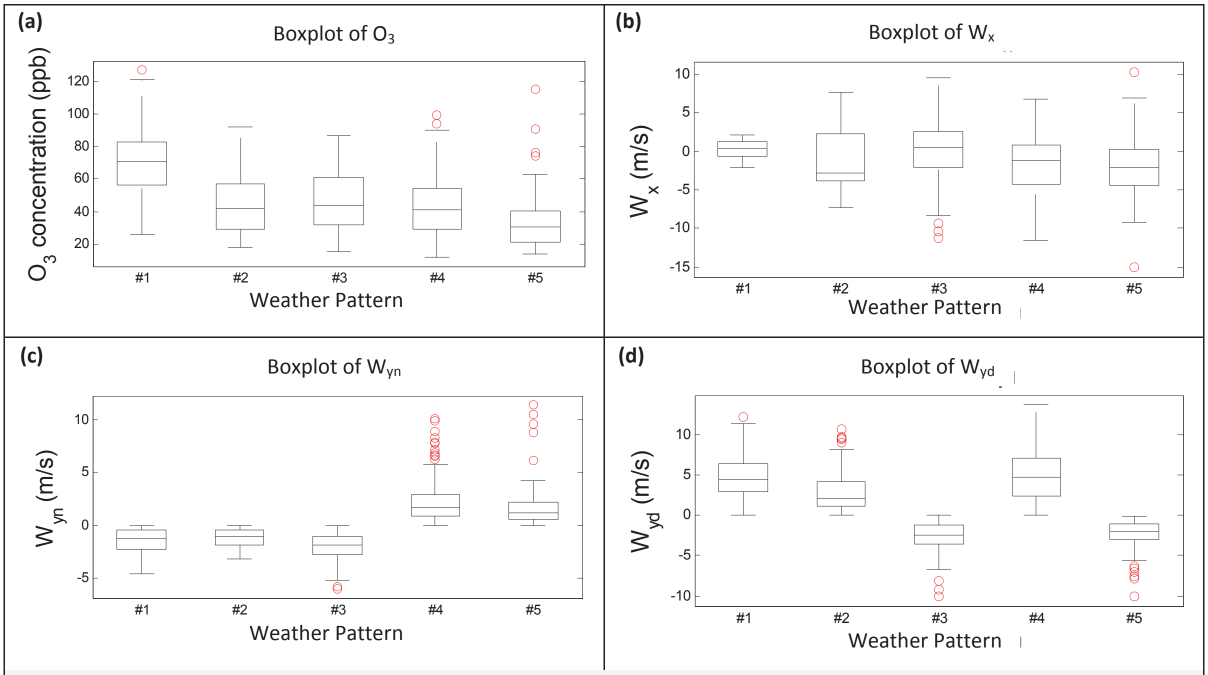

***Figure 2.*** *Boxplot of **(a)** ozone concentration, **(b)** $W_x$, **(c)** $W_{yn}$ and **(d)** $W_{yd}$ of different wind patterns. $W_x$ stands for wind speed in x direction, $W_{yn}$ stands for wind speed in y direction during nighttime and $W_{yd}$ means wind direction in y direction during day time.*

***Table 3.*** *Details of cluster results in three monitoring sites*

| Monitoring Site | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| CAMS35 | Cluster size | 188 | 94 | 113 | 424 | 104 |
| | OAD | 81 | 8 | 6 | 7 | 3 |
| | OAD percent | 43.09% | 8.51% | 5.31% | 1.65% | 2.88% |
| | Mean concentration | 69.49 | 44.27 | 46.92 | 42.61 | 34.15 |
| CAMS403 | Cluster size | 180 | 93 | 107 | 347 | 89 |
| | OAD | 42 | 3 | 0 | 3 | 2 |
| | OAD percent | 23.33% | 3.23% | 0% | 0.86% | 2.25% |
| | Mean concentration | 60.01 | 36.37 | 36.25 | 34.19 | 24.40 |
| CAMS1015 | Cluster size | 173 | 101 | 123 | 379 | 103 |
| | OAD | 60 | 5 | 2 | 6 | 2 |
| | OAD percent | 34.68% | 4.95% | 1.63% | 1.58% | 1.94% |
| | Mean concentration | 67.50 | 44.71 | 45.88 | 41.25 | 33.24 |

### 4.2. Model development

**Variable selection.** In this work, input variables used in GLMM are chosen according to the qualitative investigation of photochemical reactions occurring in boundary layer, and then the log likelihood ratio test is employed to determine their significance to the model. Because ground–level ozone is generated in the planet boundary layer by the photochemical reactions among ozone precursors under UV radiation, nitrogen oxides ($NO_X$) and highly reactive volatile organic compounds (HROCs) are considered as two main ozone precursors. As the intermediate product of photochemical chain reactions, formation and depletion rates of the current ozone are also affected by the past and current ozone concentrations. Thus, ozone concentration in the current day is usually used in prediction models. Besides air pollutants, meteorological factors also affect ozone formation significantly. According to the results by Banta et al. (1998), vector–average wind speed often accounts for more than half of the variance in daily maximum ozone values. Outdoor temperature, solar radiation and UV radiation can impact photochemical reaction rates. Usually, they are positively correlated with surface ozone concentrations while RH is negatively correlated to surface ozone concentrations (Crutzen, 1974). Correlation coefficients between ozone concentrations on the next day and its potential predictors are shown in Table 4.

**Model formulation.** Four models, which include a linear regression model, nonlinear regression model, generalized linear model and generalized linear mixed effects model, are developed to predict OADs in HGB area. Considering monitoring site CAMS35 as an illustration, linear regression model (LM) and nonlinear regression model (NLM) are built first. These models can be expressed as follows:

$$O_3(n) = 75.78 + 0.48 \times O_3(c) + 9.05 \times CO(c) - 0.20 \times NO(c) - 0.12 \times OT(n) - 0.43 \times RH(n) + 0.63 \times W_x(n) - 0.54 \times W_{yn}(n) + 0.61 \times W_{yd}(n) - 1.21 MWG(n) \quad (6)$$

$$O_3(n) = \exp(3.22 + 0.42 \times \ln(O_3(c)) - 0.0052 \times OT(n) - 0.0069 \times RH(n) + 0.14 \times SR(n) + 0.02 \times W_x(n) - 0.01 \times W_{yn}(n) + 0.01 \times W_{yd}(n) - 0.02 \times MWG(n)) \quad (7)$$

Probability density distributions of prediction error of these two models are shown in Figure 3 and Figure 4. And the corresponding histogram of the ozone values is shown in Figure 5. It is shown that ozone values are heavily non–Gaussian distributed and the OADs form a long tail in the right side. These indicate that the probability density plots of prediction errors are asymmetric and skewed. Results of Kolmogorov–Smirnov test show that both probability distribution of prediction errors are closer to gamma distribution and generalized extreme value (GEV) distribution rather than normal and log–normal distribution. Box–plot of prediction errors in different ozone levels is shown in Figure 6. It is shown that all prediction errors of ozone concentration which is above 75 ppb are positive, and the mean value is much bigger than the other three levels. When non–Gaussian distribution of modeling data is neglected, both linear and nonlinear regression models tend to underestimate ozone concentrations of OADs with larger errors than other levels.

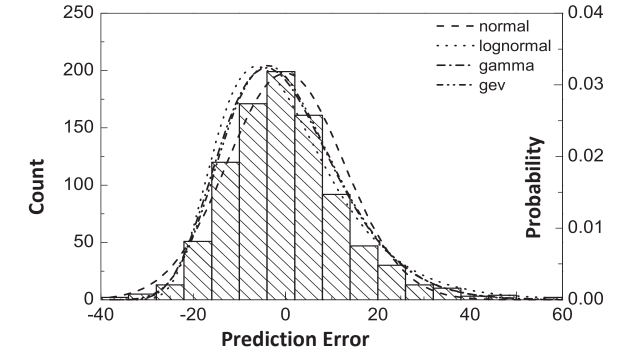

***Figure 3.*** *Probability density plot of prediction errors generated by linear regression model with data from 2003 to 2005 in HGB area.*

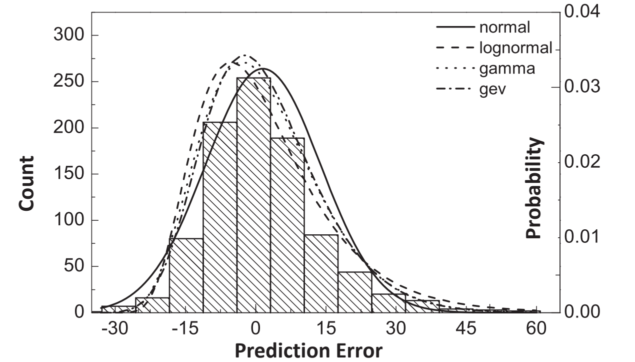

***Figure 4.*** *Probability density plot of prediction errors generated by nonlinear regression model with data from 2003 to 2005 in HGB area.*

To improve prediction accuracy of OADs, a GLM and GLMM with gamma distribution are developed. GLM is expressed as:

$$O_3(n) = (100.40 - 1.54 \times O_3(c) + 1.11 \times OT(n) + 1.17 \times RH(n) - 51.54 \times SR(n) - 6.78 \times W_x(n) + 2.32 \times W_{yn}(n) - 3.69 \times W_{yd}(n) + 6.53 \times MWG(n)) / 10\,000 \quad (8)$$

*family = gamma, link function = inverse*

The fixed effects of GLMM are the same as GLM, and its random effects should contain all variables used in fixed effects. However, results of log–likelihood ratio test suggest that OT, RH, SR and $W_x$ should be excluded from the random effects. GLMM can now be described as:

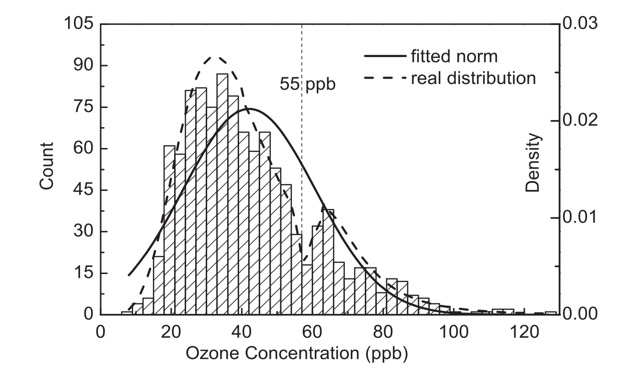

***Figure 5.*** *Histogram of the ozone value and its probability density curve.*

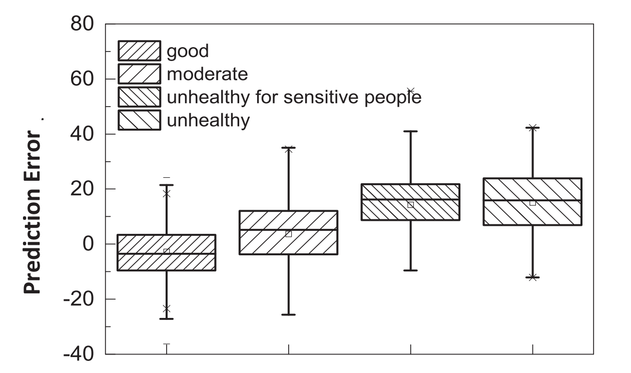

***Figure 6.*** *Box plot of prediction errors generated by linear regression model with data from 2003 to 2005 in HGB area. Where, 'good', 'moderate', 'unhealthy for sensitive people' and 'unhealthy' correspond to ozone concentration 0–30 ppb, 31–60 ppb, 61–75 ppb and above 75 ppb.*

$$\begin{aligned}
&O_3(n) = (99.7 - 1.71 \times O_3(c) + 1.11 \times OT(n) + 1.09 \times RH(n) \\
&\quad - 50.00 \times SR(n) - 6.65 \times W_x(n) + 2.32 \times W_{yn}(n) - 3.77 \\
&\quad \times W_{yd}(n) + 6.60 \times MWG(n)) / 10\,000 \\
&random = (intercept + O_3(c) + W_{yn}(n) + W_{yd}(n) \\
&\quad + MWG(n)) \mid cluster \\
&\text{family} = \text{gamma, link function} = \text{inverse}
\end{aligned} \qquad (9)$$

Coefficients of the random effects are shown in Table 5.

In GLMM, coefficient of fixed effects can describe the relationship between ozone concentrations and its predictor, which include ozone precursors and meteorological factors, on the global level; coefficients of random effects can describe relationship between ozone concentrations and its predictors in each synoptic weather pattern.

To illustrate the improvement of link function and random effects, log–likelihood ratio test is used between LM and NLM, NLM and GLM, GLM and GLMM. These results are shown in Table 6. One can observe that link function and random effects do improve performances of GLM and GLMM.

### 4.3. Statistical indicators for model performance

There are numerous statistical indicators in the air quality modeling literature for assessing the performance of air quality models. The model accuracy can be described by several performance metrics as follows:

Mean Absolute Error, $MAE = \frac{1}{N}\sum_{i=1}^{N} abs(P_i - O_i)$ (10)

Mean Bias Error, $MBE = \frac{1}{N}\sum_{i=1}^{N} (P_i - O_i)$ (11)

where, $N$ represents the sample number, $P_i$ is the predicted value and $O_i$ is the observed value.

True Positive Rate, $TPR = A/M$ (12)

False Alarm Rate, $FAR = (F–A)/(N–M)$ (13)

Success Index, $SI = TPR – FAR$ (14)

where $A$ representing the number of correctly predicted exceedances, $F$ is the number of predicted exceedances, $M$ is the number of observed exceedances, $N$ is the number of sample. Considering the different costs between false alarms and missing report, $SI$ is widely used in evaluation of surface ozone prediction models (Schlink et al., 2006). The ideal values of $MAE$ and $MBE$ is 0 ppb, ideal values of $TPR$ and $SI$ is 100% and $FAR$ is 0%, respectively.

### 4.4. Prediction results of surface ozone in HGB area

As the aim of ozone prediction model is to aid in correctly forecasting ozone action days, this paper focuses on the days in

***Table 4.*** *Correlation coefficients between surface ozone on the next day and its potential predictors, where c refers to the current day and n refers to the next day*

| Variable | $O_3$ (c) | CO (c) | NO (c) | $NO_2$ (c) | $PM_{2.5}$ OC(c) | NMHC (c) | OT (n) |
|---|---|---|---|---|---|---|---|
| CAMS35 | 0.653 | 0.280 | 0.186 | 0.471 | 0.276 | | 0.127 |
| CAMS403 | 0.627 | 0.210 | –0.402 | 0.349 | 0.349 | | 0.177 |
| CAMS1015 | 0.626 | | | 0.300 | | 0.368 | 0.159 |
| | RH(n) | SR(n) | UR(n) | $W_x$(n) | $W_{yn}$(n) | $W_{yd}$(n) | MWG(n) |
| CAMS35 | –0.509 | 0.400 | | 0.429 | –0.292 | 0.155 | –0.392 |
| CAMS403 | –0.414 | 0.440 | 0.229 | 0.491 | –0.139 | 0.111 | –0.501 |
| CAMS1015 | | 0.426 | | 0.410 | –0.165 | –0.075 | –0.508 |

***Table 5.*** *Random effect coefficients of the GLMM in Equation (9)*

| Cluster | Intercept | $O_3$ (c) | $W_{yn}$ (n) | $W_{yd}$ (n) | MWG (n) |
|---|---|---|---|---|---|
| 1 | $–3.61 \times 10^{-4}$ | $–0.12 \times 10^{-4}$ | $0.20 \times 10^{-4}$ | $–0.31 \times 10^{-4}$ | $0.07 \times 10^{-4}$ |
| 2 | $–0.79 \times 10^{-4}$ | $–0.09 \times 10^{-4}$ | $0.24 \times 10^{-4}$ | $–0.27 \times 10^{-4}$ | $0.09 \times 10^{-4}$ |
| 3 | $–0.57 \times 10^{-4}$ | $–0.09 \times 10^{-4}$ | $0.27 \times 10^{-4}$ | $–0.18 \times 10^{-4}$ | $0.07 \times 10^{-4}$ |
| 4 | $0.00 \times 10^{-4}$ | $–0.00 \times 10^{-4}$ | $0.31 \times 10^{-4}$ | $–0.18 \times 10^{-4}$ | $0.07 \times 10^{-4}$ |
| 5 | $–0.00 \times 10^{-4}$ | $–0.02 \times 10^{-4}$ | $0.25 \times 10^{-4}$ | $–0.22 \times 10^{-4}$ | $0.05 \times 10^{-4}$ |

***Table 6.*** *Details of log–likelihood ratio test between different models*

| Model | #DF | Loglike | Df | Chisq | Critical Value (0.01) |
|---|---|---|---|---|---|
| LM | 11 | –3 649.7 | | | |
| NLM | 10 | –3 638.1 | 1 | 21.24 | 6.63 |
| GLM | 10 | –3 600.7 | 0 | 74.82 | NaN |
| GLMM | 15 | –3 561.0 | –5 | 79.4 | 15.08 |

which maximum 8 h–average ozone concentration is 75 ppb and above. Percentages of ozone action days at CAMS35, CAMS403 and CAMS1015 are 11.37%, 6.37% and 8.53%, respectively. Prediction models are built with data from 2003 to 2005. Data from 2006 to 2007 are used to test these models. During model development, a *J*–fold cross–validation method is used to determine model parameters. In the cross–validation step, data from two years are used as training data and data from the other one year is used for validation. Data from each year from 2003 to 2005 are used as training data and validation data. To illustrate the superiority of this proposed cluster based GLMM, MLPs and SVMs with complex structures also used to predict ozone exceedances in HGB area. The MLP consists an input layer with 9 input nodes, one hidden layer consists of 4 nodes, and an output layer with a single output node. The hidden layer nodes use a sigmoid transfer function to generate outputs. The structure of the SVM model is the same as the MLP used. The kernel function used in SVM is also sigmoid. Both MLP and SVM are applied using Matlab 2009b. A comparison of model performance during modeling step is shown in Table 7.

Table 7 shows that TPR of both GLM and GLMM are improved significantly compared to LM. Although MAE of GLMM is not improved notably, MAE of OADs generated by GLMM is decreased by an average of 21.83%, compared to LMs. Comparisons between $MAE_{OAD}$ and $MBE_{OAD}$ of LM show that LMs are prone to underestimate high ozone concentrations systematically. And this situation has been improved in GLMMs. Though MLP and SVM can model surface ozone concentrations with much lower MAEs, they also underestimate ozone levels of OADs systematically which lead to lower SIs than GLMM. Validation results show that cluster based GLMM can predict 80.86% OADs correctly, and the average FAR is 3.46%.

When GLMMs are determined, data of ozone season from 2006 to 2007 are used to test the model. There are 26, 6 and 18 OADs at CAMS35, CAMS403 and CAMS1015, respectively. Compared to the modeling data, OAD percentage in the test data is much less. Test performances of these models are shown in Table 8. Table 8 shows that GLMMs can predict 86.18% OADs correctly during test step, which is increased by 4.37% compared to modeling step. While, MLP and SVM can predict 61.26% and 66.95% OADs correctly, which is 28.92% and 22.31% lower that of GLMMs. Although these GLMMs are robust to new data, they tend to have higher FAR during the test step. Under the same meteorological conditions, surface ozone concentrations during 2006 and 2007 are lower than the modeling period. This suggests that the parameters of GLMMs should be updated in time. Compared to the model used by TCEQ currently, TPR of OADs is increased by 19% during test step.

## 5. Conclusions

GLMMs based on cluster analysis are developed to improve TPR of OADs in HGB area. A two–step clustering method is used to identify weather patterns which are most likely to be coincident with OADs. Then, GLMMs are used to make predictions. With the link function, GLMMs are able to model ozone action days which locate the right tail of probability density plot; with the random effects, GLMMs are able to model the differences of meteorological effects on ozone formation and dispersion. Compared to linear regression and generalized linear models, GLMM can improve both prediction accuracy and TPR of ozone action days significantly. The model proposed in this paper also outperforms the current prediction model used by TCEQ at the selected monitoring sites. The test results show that these GLMMs are robust to new data. Compared to the MLPs and SVMs, cluster based GLMM can capture the wind field features which are favorable for surface ozone formation and accumulation and build explanatory models between surface ozone levels and the influence factors.

Although the proposed model can improve TPR of ozone action days to some degree, there is room for further improvement in TPR. As a data–driven model, the prediction accuracy of this prediction model depends on the quality of monitored ozone precursor data and meteorological data. Because air pollutant data in the next day is not available, this model is unable to reflect the effects of emission events happened in the next day. Thus it may fail to report ozone action days caused by emission events.

***Table 7.*** *Comparison of five model performances at three sites during training step*

| Sites | Model | MAE | $MAE_{OAD}$ | $MBE_{OAD}$ | TPR | FAR | SI |
|---|---|---|---|---|---|---|---|
| CAMS35 | LM | 9.30 | 17.59 | 17.01 | 34.29% | 1.35% | 32.94% |
| | GLM | 9.65 | 18.11 | 11.90 | 45.71% | 3.27% | 42.44% |
| | GLMM | 8.51 | 13.75 | 8.81 | 85.71% | 3.31% | 82.40% |
| | MLP | 5.77 | 11.70 | 10.51 | 68.57% | 3.27% | 65.30% |
| | SVM | 5.41 | 10.90 | 10.14 | 73.33% | 1.96% | 71.37% |
| CAMS403 | LM | 9.34 | 23.32 | 22.95 | 14.00% | 0.52% | 13.48% |
| | GLM | 10.07 | 25.77 | 23.99 | 36.00% | 2.35% | 33.65% |
| | GLMM | 9.65 | 19.12 | 14.45 | 82.00% | 3.31% | 78.69% |
| | MLP | 5.13 | 15.86 | 14.71 | 72.00% | 2.94% | 69.06% |
| | SVM | 4.98 | 15.14 | 13.72 | 78.00% | 3.43% | 74.57% |
| CAMS1015 | LM | 8.70 | 18.75 | 18.11 | 14.67% | 1.25% | 13.42% |
| | GLM | 8.81 | 17.81 | 14.97 | 36.00% | 2.84% | 33.16% |
| | GLMM | 8.74 | 11.17 | 6.29 | 80.00% | 3.53% | 76.47% |
| | MLP | 4.77 | 9.71 | 9.05 | 68.00% | 2.84% | 65.16% |
| | SVM | 4.51 | 9.19 | 8.90 | 73.33% | 3.30% | 70.03% |

***Table 8.*** *Comparison of five model performances at three sites during test step*

| Sites | Model | MAE | $MAE_{OAD}$ | $MBE_{OAD}$ | TPR | FAR | SI |
|---|---|---|---|---|---|---|---|
| CAMS35 | LM | 10.04 | 18.71 | 17.79 | 38.46% | 1.93% | 36.53% |
| | GLM | 10.17 | 19.22 | 12.79 | 57.69% | 3.87% | 53.82% |
| | GLMM | 9.77 | 14.19 | 9.26 | 80.76% | 3.79% | 76.97% |
| | MLP | 7.40 | 12.11 | 11.42 | 61.54% | 3.59% | 57.95% |
| | SVM | 7.07 | 11.53 | 10.70 | 73.08% | 2.49% | 70.69% |
| CAMS403 | LM | 9.89 | 24.70 | 24.01 | 16.67% | 0.95% | 15.72% |
| | GLM | 10.94 | 24.31 | 22.71 | 33.33% | 2.61% | 30.72% |
| | GLMM | 9.82 | 20.57 | 14.70 | 100.00% | 3.50% | 96.50% |
| | MLP | 6.79 | 16.26 | 15.22 | 66.67% | 3.55% | 63.12% |
| | SVM | 6.47 | 15.90 | 14.12 | 66.67% | 2.61% | 64.06% |
| CAMS1015 | LM | 9.43 | 17.69 | 17.14 | 16.67% | 1.45% | 15.22% |
| | GLM | 9.77 | 18.15 | 16.42 | 44.44% | 3.49% | 40.95% |
| | GLMM | 9.01 | 11.17 | 6.29 | 77.78% | 3.71% | 74.07% |
| | MLP | 8.41 | 12.77 | 11.09 | 55.56% | 4.10% | 51.46% |
| | SVM | 5.75 | 11.03 | 10.39 | 61.11% | 3.30% | 57.81% |

## Acknowledgments

This work was supported by Fundamental Research Funds for the Central Universities (Item No. SWU114021) and partially supported by grants from the National Natural Science Foundation of China (Item No. 51203008). Also, the authors would like to thank anonymous reviewers and editors for their useful comments which have helped to improve the quality of this manuscript.

## References

Banta, R.M., Senff, C.J., White, A.B., Trainer, M., McNider, R.T., Valente, R.J., Mayor, S.D., Alvarez, R.J., Hardesty, R.M., Parrish, D., Fehsenfeld, F.C., 1998. Daytime buildup and nighttime transport of urban ozone in the boundary layer during a stagnation episode. *Journal of Geophysical Research–Atmospheres* 103, 22519–22544.

Beaver, S., Palazoglu, A., 2006a. A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. *Atmospheric Environment* 40, 713–725.

Beaver, S., Palazoglu, A., 2006b. Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality. *Journal of Applied Meteorology and Climatology* 45, 1710–1726.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.S.S., 2009. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 127–135.

Byun, D.W., Kim, S.T., Kim, S.B., 2007. Evaluation of air quality models for the simulation of a high ozone episode in the Houston metropolitan area. *Atmospheric Environment* 41, 837–853.

Cobourn, W.G., Hubbard, M.C., 1999. An enhanced ozone forecasting model using air mass trajectory analysis. *Atmospheric Environment* 33, 4663–4674.

Crutzen, P.J., 1974. Photochemical reactions initiated by and influencing ozone in unpolluted tropospheric air. *Tellus* 26, 47–57.

Darby, L.S., 2005. Cluster analysis of surface winds in Houston, Texas, and the impact of wind patterns on ozone. *Journal of Applied Meteorology* 44, 1788–1806.

Davis, J.M., Speckman, P., 1999. A model for predicting maximum and 8 h average ozone in Houston. *Atmospheric Environment* 33, 2487–2500.

Davis, J.M., Eder, B.K., Nychka, D., Yang, Q., 1998. Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Environment* 32, 2505–2520.

Done, J., Davis, C.A., Weisman, M., 2004. The next generation of NWP: Explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmospheric Science Letters* 5, 110–117.

Draxler, R.R., 2000. Meteorological factors of ozone predictability at Houston, Texas. *Journal of the Air & Waste Management Association* 50, 259–271.

Emmons, L.K., Walters, S., Hess, P.G., Lamarque, J.F., Pfister, G.G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S.L., Kloster, S., 2010. Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART–4). *Geoscientific Model Development* 3, 43–67.

Fast, J.D., Gustafson, W.I., Easter, R.C., Zaveri, R.A., Barnard, J.C., Chapman, E.G., Grell, G.A., Peckham, S.E., 2006. Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of Houston using a fully coupled meteorology–chemistry–aerosol model. *Journal of Geophysical Research–Atmospheres* 111, art. no. D21305.

Fauroux, B., Sampil, M., Quenel, P., Lemoullec, Y., 2000. Ozone: A trigger for hospital pediatric asthma emergency room visits. *Pediatric Pulmonology* 30, 41–46.

Feng, Y., Zhang, W.F., Sun, D.Z., Zhang, L.Q., 2011. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. *Atmospheric Environment* 45, 1979–1985.

Groves, A.M., Gow, A.J., Massa, C.B., Laskin, J.D., Laskin, D.L., 2012. Prolonged injury and altered lung function after ozone inhalation in mice with chronic lung inflammation. *American Journal of Respiratory Cell and Molecular Biology* 47, 776–783.

Henze, D.K., Hakami, A., Seinfeld, J.H., 2007. Development of the adjoint of GEOS–Chem. *Atmospheric Chemistry and Physics* 7, 2413–2433.

Jenkin, M.E., Clemitshaw, K.C., 2000. Ozone and other secondary photochemical pollutants: Chemical processes governing their formation in the planetary boundary layer. *Atmospheric Environment* 34, 2499–2527.

Ji, M., Cohan, D.S., Bell, M.L., 2011. Meta–analysis of the association between short–term exposure to ambient ozone and respiratory hospital admissions. *Environmental Research Letters* 6, art. no. 024006.

Jiang, G.F., Fast, J.D., 2004. Modeling the effects of VOC and $NO_x$ emission sources on ozone formation in Houston during the TexAQS 2000 field campaign. *Atmospheric Environment* 38, 5071–5085.

Misenis, C., Zhang, Y., 2010. An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options. *Atmospheric Research* 97, 315–334.

Neidell, M., 2010. Air quality warnings and outdoor activities: Evidence from Southern California using a regression discontinuity design. *Journal of Epidemiology and Community Health* 64, 921–926.

Prybutok, V.R., Yi, J.S., Mitchell, D., 2000. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research* 122, 31–40.

Psichogios, D.C., Ungar, L.H., 1992. A hybrid neural network–first principles approach to process modeling. *AIChE Journal* 38, 1499–1511.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E., 2006. Statistical models to assess the health effects and to forecast ground–level ozone. *Environmental Modelling & Software* 21, 547–558.

Schneider, T., 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14, 853–871.

Sfetsos, A., Siriopoulos, C., 2004. Time series forecasting with a hybrid clustering scheme and pattern recognition. *IEEE Transactions on Systems Man and Cybernetics Part A–Systems and Humans* 34, 399–405.

Sfetsos, A., Vlachogiannis, D., Gounaris, N., 2013. An investigation of the factors affecting the ozone concentrations in an urban environment. *Atmospheric and Climate Sciences* 3, 11–17.

Sun, W., Zhang, H., Palazoglu, A., 2013. Prediction of 8 h–average ozone concentration using a supervised hidden Markov model combined with generalized linear models. *Atmospheric Environment* 81, 199–208.

Tesche, T.W., Morris, R., Tonnesen, G., McNally, D., Boylan, J., Brewer, P., 2006. CMAQ/CAMx annual 2002 performance evaluation over the eastern US. *Atmospheric Environment* 40, 4906–4919.

Tolbert, P.E., Mulholland, J.A., MacIntosh, D.L., Xu, F., Daniels, D., Devine, O.J., Carlin, B.P., Klein, M., Dorley, J., Butler, A.J., Nordenberg, D.F., Frumkin, H., Ryan, P.B., White, M.C., 2000. Air quality and pediatric emergency room visits for asthma in Atlanta, Georgia. *American Journal of Epidemiology* 151, 798–810.

Vizuete, W., Kim, B.U., Jeffries, H., Kimura, Y., Allen, D.T., Kioumourtzoglou, M.A., Biton, L., Henderson, B., 2008. Modeling ozone formation from industrial emission events in Houston, Texas. *Atmospheric Environment* 42, 7641–7650.

Zahedi, G., Saba, S., Elkamel, A., Bahadori, A., 2014. Ozone pollution prediction around industrial areas using fuzzy neural network approach. *CLEAN–Soil Air Water* 42, 871–879.

Zhang, K., Fan, W., 2008. Forecasting skewed biased stochastic ozone days: Analyses, solutions and beyond. *Knowledge and Information Systems* 14, 299–326.

Zhang, H., Palazoglu, A., Zhang, X.Y., Zhang, W.D., Zhao, Z.M., Sun, W., Liu, S.W., 2014. Prediction of surface ozone exceedance days using PCA with a non–parametric $T^2$ control limit. *Chemometrics and Intelligent Laboratory Systems* 133, 42–48.

Zhang, F.Q., Bei, N.F., Nielsen–Gammon, J.W., Li, G.H., Zhang, R.Y., Stuart, A., Aksoy, A., 2007. Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts. *Journal of Geophysical Research–Atmospheres* 112, art. no. D04304.