Journal of the Air & Waste Management Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uawm20

Forecasting Peak Daily Ozone Levels—I. A Regression with Time Series Errors Model Having a Principal Component Trigger to Fit 1991 Ozone Levels

Pao-Wen Grace Liu [a] & Richard Johnson [b]

[a] Bureau of Air Management, Wisconsin Department of Natural Resources, Madison, Wisconsin, USA

[b] Department of Statistics, University of Wisconsin, Madison, Wisconsin, USA

Published online: 27 Dec 2011.

TECHNICAL PAPER

ISSN 1047-3289 *J. Air & Waste Manage. Assoc.* **52**:1064-1074

Forecasting Peak Daily Ozone Levels—I. A Regression with Time Series Errors Model Having a Principal Component Trigger to Fit 1991 Ozone Levels

Pao-Wen Grace Liu
Bureau of Air Management, Wisconsin Department of Natural Resources, Madison, Wisconsin

Richard Johnson
Department of Statistics, University of Wisconsin, Madison, Wisconsin

ABSTRACT
This research was motivated by the need to warn the population of Milwaukee, WI, on high-ozone days. A statistical model for the peak daily 1-hr ozone level is proposed. A Regression with Time Series Errors (RTSE) model, which includes a principal component (PC) trigger, is the basis for forecasting the peak daily 1-hr ozone level.

The RTSE model, with a PC trigger, is first employed to estimate daily peak ozone measured at the University of Wisconsin, Milwaukee-North (UWM-N), during the 1991 ozone season. The RTSE model uses peak daily temperature, morning vector average wind direction, and the PC trigger as predictor variables. The PC trigger was designed to summarize atmospheric circumstances when peak ozone was greater than 100 parts per billion (ppb). It is verified that the RTSE model, with a PC trigger, significantly improves the prediction of peak daily ozone, particularly peak ozone greater than 100 ppb. In comparison with the RTSE model without the PC trigger, the RTSE model with a PC trigger raised the R^2 from 0.680 to 0.809.[1] It is suggested that the RTSE model, with the PC trigger, is an adequate statistical model that has the potential for real-time ozone forecasting.

IMPLICATIONS
Once this statistical model is shown to adequately fit historical data, it can be used to make real-time forecasts of peak ozone. Milwaukee has been designated as a severe 1-hr ozone nonattainment area for approximately 10 years. Early warning on days with peak high ozone can help people avoid exposures to high ozone, which causes or exacerbates human respiratory problems. We developed an ozone model, RTSE and a PC trigger, that solves some problems of underprediction of high ozone and also provides the potential to make actual real-time ozone forecasts.

INTRODUCTION
The purpose of this research is to design, test, and implement a technique to forecast next-day peak daily 1-hr ambient ozone levels. A pilot model is developed. The application of this pilot model to real-time forecasting in Milwaukee, WI, will be detailed in a companion paper.[2]

Ozone in the stratosphere, about 15–55 km altitude, plays a critical role in protecting humans from harmful UV radiation. However, ozone in the troposphere between the Earth's surface and 10–15 km altitude is a harmful pollutant that causes human health problems. The ground-level or ambient ozone is known as a secondary pollutant and is produced from photochemical reactions under certain meteorological conditions. Anthropogenic and biogenic volatile organic compounds (VOCs) and nitrogen oxides (NO_x) are found to be major precursors of ozone pollution.

During the past several decades, ambient ozone has been a serious environmental problem in Milwaukee. The Federal Clean Air Act Amendments of 1977 and 1990 resulted in the designation of southeastern Wisconsin as a severe 1-hr ozone nonattainment area. The Wisconsin Department of Natural Resources (WDNR) submitted ozone State Implementation Plans (SIPs) in 1979, 1983, and 2000 to demonstrate attainment of the 1-hr ozone standard.[3] Reformulated gasoline, improved motor vehicle emissions control, and control of industrial service have been very effective in reducing ozone concentration in eastern Wisconsin. Additionally, voluntary measures such as Ozone Action Days (OADs), cooperatively conducted by the Wisconsin State government and industries, have helped reduce emissions of ozone precursors. The downward trend of the temperature-adjusted ozone over the period 1980–1995 verifies the effectiveness of the ozone control programs.[4] The number of days of violating the 1-hr ozone standard also has been gradually reduced.[4] However, Milwaukee County is still designated as one of the six severe 1-hr ozone nonattainment counties (see Figure 1).

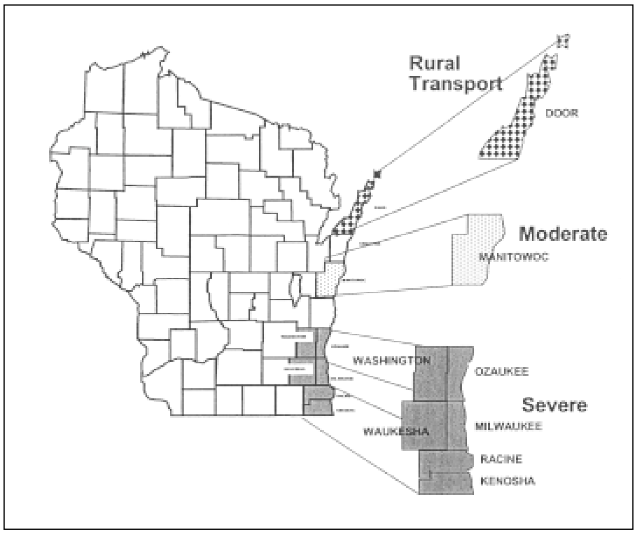

Figure 1. Ozone 1-hr nonattainment areas in southeastern Wisconsin in 1998.

Many different techniques have been used to analyze ozone data during the past 20 years. A common problem is that these errors tend to be much larger for high ozone, and those high concentrations are underpredicted.[5-8] Also, these models were verified with archived data only. Most of the forecasts can be fairly accurate during the days that ozone did not exceed the 1-hr standard and temperatures were not dramatically high.[6,9,10] When these models were constructed, they were based on entire seasons, which contain mostly average-ozone days. Consequently, those models failed to predict the high-ozone days that are usually only ~5% of the entire season.[11] Therefore, it is essential to develop a model with improved accuracy at the higher ozone levels that can be used to forecast ozone in real time.

The objective of this research is to build a pilot ozone model that can successfully improve the prediction of high ozone concentrations and that has the potential to make real-time forecasts. A special case of the Box-Jenkins transfer function model, a Regression with Time Series Errors (RTSE) model, is employed along with a principal component (PC) analysis to construct the pilot ozone model.[12,13] The purpose of the PC analysis is to create a trigger for effectively estimating high ozone concentrations. The proposed ozone model is called the RTSE model with a PC trigger.

STATEMENT OF THE PROBLEM

Figure 1 indicates the severity of the ozone exceedance problem in Wisconsin. Six counties, including Milwaukee, are designated as severe nonattainment counties and the two other counties are designated moderate and rural transport nonattainment. Though during the past 10 years, the 1-hr ozone nonattainment area has been reduced from 11 to eight counties, Milwaukee County is still designated as a severe 1-hr ozone nonattainment area.

Wisconsin has developed a subjective forecast system operated by the state meteorologists. Beginning in the summer of 1995, southeastern Wisconsin has been participating in a voluntary effort to help reduce ground-level ozone concentrations in the southern Lake Michigan region. The Lake Michigan Ozone Region includes southeastern Wisconsin, northeastern Illinois, northwestern Indiana, and western Michigan (see Figure 2). In 1995, the WDNR joined with the public media and private enterprises in beginning a program of prior-day notification for those days in eastern Wisconsin that are forecast to have weather conditions favorable for the production of high ozone. Such a day is called an OAD. However, the OAD program is established based on a qualitative description of the weather, and no specific ozone level or range of values is produced by this forecast. The goal of the OAD program is to be able to notify the public in time to reduce traffic emission and to avoid ozone exposure.

Photochemical models, such as the urban airshed model (UAM), are difficult to calibrate and require intensive computer resources.[14] Simulation of a single ozone event requires a considerable amount of data input (including emission, meteorological, and chemistry data), and it is time-consuming.[5,15] Particularly, the UAM model was designed for regulatory purposes to validate strategies of reducing ozone precursors. Chang and Cardelino[16] conclude that the numerical processing barrier of applying UAM to forecast peak ozone has been removed by using low-cost and powerful workstations. However, in Wisconsin, the UAM was used only to demonstrate how to attain the 1-hr ozone standard in the Lake Michigan area—not to make immediate forecasts.[17] WDNR research

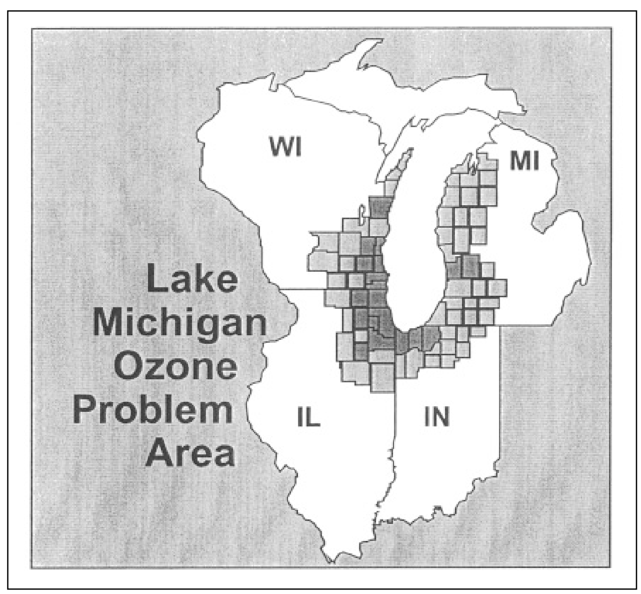

Figure 2. The Lake Michigan Ozone Study area. Counties shaded dark are in nonattainment areas in Wisconsin, Illinois, Indiana, and Michigan of the 1-hr NAAQS for ozone, 1991.

emphasizes that the UAM systematically underpredicts high ozone values.[18]

Mathematical/statistical models, such as regression models, have difficulty predicting extreme ozone values.[5-8] Underpredicting and overpredicting frequently are observed among the ozone modeling studies. Most ozone modeling focuses on estimating the long-term ozone trend or implementing emission-control policy.[9,10,19,20]

A frequently observed problem in regression analysis is that some of the employed regression models seriously violate a required assumption. The models were designed without noticing, mentioning, or fulfilling the fundamental assumption that the errors in the model have to be independent and identically distributed (i.i.d. assumption). In addition, most of the regression studies used all of the ozone data from an entire season to construct their forecast models. During an entire ozone season, the peak daily ozone values vary greatly from 30 to 200 parts per billion (ppb), but the extremely high ozone episodes appear to be always less than 5% of the entire season.[11] Only three studies improved their forecasts by using data stratified according to ozone level.[7,8,21]

Box-Jenkins[12] Auto-Regressive Integrated Moving Average (ARIMA) time-series models provide another method to predict series of daily ozone. Time-series models relax the assumption of independence and are consistent with the nature of serially correlated air pollutants. Though the problems of underpredictions and overpredictions are still observed, some research has emphasized that Box-Jenkins multivariate time-series models are superior to their univariate time-series model in predicting ozone concentrations.[22-24]

A rather severe weakness in existing forecasting procedures has been indicated. As a result, the RTSE model with a PC trigger was developed to solve the significant underestimation of high ozone concentrations. This method also is proposed because of its capability of quantitative forecasting and its inexpensive operations.

METHODS

Forecasting systems combining a nonlinear regression with an ozone-conducive criteria have been recommended to improve forecasting accuracy.[8,21,25] Our research will use Box-Jenkins multivariate time-series models in combination with PC analysis.[12,13] The special form of the Box-Jenkins transfer function model, an RTSE model, is combined with PC analysis. In our proposed RTSE model, the PC trigger, in addition to meteorological and NO_x predictors, is one of the predictors. Particularly, the PC trigger in the RTSE model is intended to predict ozone above 100 ppb. For regular ozone days, the RTSE model forecasts the ozone values as a baseline, and the PC trigger is turned off. For high-ozone days, the PC trigger is turned on to increase the prediction of peak ozone above the baseline prediction.

RTSE Model

An RTSE model is advantageous for two reasons. First, the time dependence of ozone and meteorological variables are modeled. Air pollution, particularly ozone, is highly correlated over time.[26,27] Second, an RTSE model can incorporate reduced dimensional summaries of multivariate meteorological variables. Indeed, both meteorology variables and ozone precursors have a great influence on ozone formation and both have been considered in ozone modeling.[8,9,14,28-30]

The command proc arima in the SAS statistical software package was used to fit and analyze our time-series models to build an RTSE model.[31] In the usual notation, let B be the backward shift operator so that $\phi_1 BY_t = \phi_1 Y_{t-1}$ and $\phi_2 B^2 Y_t = \phi_2 Y_{t-2}$. The Greek letters denote parameters to be estimated. One example of an RTSE model is

$$\begin{aligned} Y_t &= \nu_1 X_{1,t} + \nu_2 X_{2,t} + \frac{a_t}{(1-\phi_1 B - \phi_2 B^2)} \\ &= \nu_1 X_{1,t} + \nu_2 X_{2,t} + N_t \end{aligned} \tag{1}$$

where Y_t stands for the response variable (ozone), and the predictor variables $X_{1,t}$ and $X_{2,t}$ could be temperature and wind speed. The noise N_t arises from an innovation series a_t, which consists of a sequence of i.i.d. assumption random shocks. The noise term N_t in eq 1 follows an AR(2) time-series model, as opposed to classic regression models, as detailed in the appendix.[12]

The process of developing an RTSE model is presented as a flow chart in Figure 3. The first three major steps are sequentially displayed in the corresponding box.

(1) Identify the structure of N_t (eq 1);
(2) Conduct a regression with time series errors model. The N_t could be similar to that in eq 1; and
(3) Append future inputs to the original data set to make forecasts. This is a crucial step in making the RTSE flexible enough to incorporate updated values for the predictor variables obtained from other forecasting methods.

Identify the Structure of Noise N_t. As a benchmark, the ozone series is first fit with respect to its own history and without using any predictor variables. Then it can be determined if the inclusion of any of the candidate predictors listed in Table 1 results in substantial improvements. According to Milionis and Davies,[26] with no predictor variable, the model structure of the noise N_t should resemble that of the response variable. The stochastic structure of N_t quantifies the amount of dependence in the ozone series. If the predictor variables ($X_{1,t}$ and $X_{2,t}$) in eq 1 are not important, the ARIMA model in eq 1 reduces to N_t only.

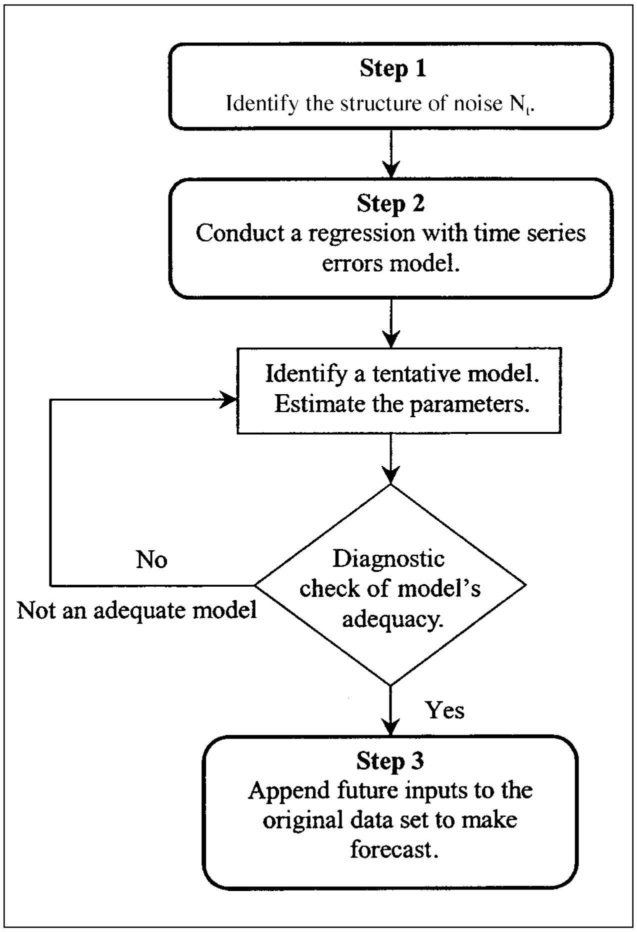

Figure 3. Process of building an RTSE model.

While this noise could be an initial approximation, it is likely to change when predictor variables are introduced. After introducing predictor variables, N_t must be updated.

A good ARIMA model is typically statistically adequate and parsimonious for a fitting given realization of a time series.[31] The principal of parsimony is characterized by fitting a set of data with the smallest number of estimated parameters. In addition to being parsimonious, a good ARIMA model has to be stationary. A stationary series has mean, variance, and autocorrelation coefficients that are essentially constant through time.[12,32] Nonstationary data often can be transformed to stationary data by using differencing or transformations. Transformation, such as taking a logarithm, can transform data to have constant variance over time.

Conduct an RTSE Model. The three stages of building an RTSE model are similar to those of developing an ARIMA model: identification, estimation, and diagnostic checking. The identification and estimation stages may refer to Step 2 in Figure 3, and the diagnostic-checking stage refers to the decision process "check model's adequacy" in Figure 3. This step resulted in a tentative model by determining its significant parameters. The parameters are estimated in an iterative way with software that simultaneously estimates both the parameters of the regression and the N_t to achieve a minimum residual variance (or residual mean square). SAS software can achieve this purpose and was employed in this study.[31,33]

The diamond shape in Figure 3 indicates that diagnostic checking is a decision process that decides whether to further revise the model. If the tentative model is not adequate, another model structure should be identified. This process should be iterated until an adequate model is achieved. Other predictor values and other noise structures can be examined. The Ljung-Box statistics and residual autocorrelation functions are employed primarily for checking the adequacy of any candidate model.[12,32,34]

Append Future Inputs to the Original Data Set to Make Forecasts. When making forecasts with the RTSE model, the future values of predictors, such as the forecast temperature and forecast wind speed, must be obtained from outside sources. Future values of predictors need to be

Table 1. Response variable and predictor variables in the ozone model development.

Response Variable and Candidate Predictor Variables		Notation	Significant Predictor Variable
Ozone	Natural logged peak daily 1-hr ozone concentration (ppb)	*lnO3*	
Temperature	Peak daily 1-hr temperature (°F)	*Tp*	*Tp*
Solar radiation	Average solar radiation from 5:00 a.m. to 8:00 p.m. (MJ/m^2)	*SR*	
Dew-point temperature	24-hr averaged dew-point temperature (°F)	*Dpt*	
NO_x	Averaged NO_x from 6:00 a.m. to 9:00 a.m. (ppb)	NO_x	
Morning wind speed	Vector averaged wind speed from 5:00 a.m. to 10:00 a.m. (mph)	*WS510*	
Noon-afternoon wind speed	Vector averaged wind speed from 10:00 a.m. to 7:00 p.m. (mph)	*WS1019*	
Morning wind direction	Vector averaged wind direction from 5:00 a.m. to 10:00 a.m. (degree)	*WD510*	*WD510*
Noon-afternoon wind direction	Vector averaged wind direction from 10:00 a.m. to 7:00 p.m. (degree)	*WD1019*	
PC	0 or $0.033Tp - 0.052NO_x + 0.888WD510 + 0.456WD1019$	*PC*	*PC*

appended to the original data set. The method for how the future values will be appended to the original data set and incorporated into the forecasts is detailed in a companion paper.[2]

PC Analysis

To improve the prediction of extremely high ozone concentration (ozone greater than 100 ppb), a PC analysis is performed. The PC in this study is a linear combination of the weather and NO_x variables during high-ozone days, and it is anticipated to demonstrate the levels of NO_x and weather conditions associated with high ozone. The resultant PC is employed as a trigger and actually is included as one of the predictors in the RTSE model. For days when observed ozone is greater than 100 ppb, the PC trigger is turned on. This trigger is calculated according to eq 2. For days with ozone below 100 ppb, the PC trigger is turned off and set to 0.

Theory. The general objectives of a PC analysis are data reduction and data interpretation.[13] For example, starting with a large number of *p* variables, PC analysis reduces to *k* variables, which can reproduce almost as much as the original *p* variables. The original data set consisting of *n* measurements on *p* variables is reduced to summary data consisting of *n* measurements on *k* principal components. Each PC is a linear combination of the original *p* variables. The first of the *k* components reproduces the most variability and is called the first PC. The first PC of NO_x and the meteorological variables were obtained using sas princorm, which is the command in the SAS statistical software package.[34]

PC Development. Two PC analyses were conducted from two different databases. The first candidate PC was computed from days that had ozone above 100 ppb from 1987–1998. The second candidate was generated from those days with ozone levels greater than 100 ppb during 1993–1998. The high-ozone days usually were separated in time, so we treated the day-to-day variation in their atmospheric conditions as independent. From 1983 to 1990, improved emission control measures resulted in a steady decrease in ozone precursor emissions.[35] This steady decrease continued throughout the 1990s largely as a result of new control measures implemented under the 1990 amendments to the federal Clean Air Act. The 1987–1998 database presents a variety of environments for high-ozone days, compared with the 1993–1998 database, in which many of the emission controls have already been implemented. Also, the 1993–1998 database excludes the extreme drought summer of 1988 and the associated anomalous weather conditions.[35]

The comparison of the two candidate first PCs is listed in Table 2. The notation for the variables is given in Table 1. Because PC II explained more variance than PC I, PC II was selected for the forecast model in this study. PC II explained 85.9% of the variability of the weather and NO_x during the days when ozone was above 100 ppb from 1993–1998, while PC I only explained 77% of the variability for the high-ozone days during 1987–1998. Because explaining maximum variance in one data set is no guarantee that the component will be the best predictor, a couple of alternative PCs were also investigated. The first PC of PC II was replaced with the second PC. There was a slight deterioration in predicting ability. We also tried obtaining PC II with temperature and morning vector average wind direction (WD510) excluded. However, our visual inspection and their larger estimated variance both indicate that neither the first nor the second PC of the changed PC II can predict high-ozone days well.

For parsimony, PC II was simplified as eq 2. It still explains the variance as well as that in Table 2. Therefore, PC II is determined to be the PC trigger in the proposed RTSE model.

$$PC_t = 0.033\ Tp_t - 0.052\ NO_{xt} + 0.888\ WD510_t + 0.456\ WD1019_t \quad (2)$$

Table 2. PC analysis.

Analysis	Database Year Domain	Proportion of Explained Variance
Analysis I[a]		
$PC\ I = 0.025\ Tp - 0.058\ NO_x - 0.004\ Dpt + 0.890\ WD510 + 0.014\ WS510 + 0.450\ WD1019$	1987–1998	0.770
Analysis II		
$PC\ II = 0.033\ Tp - 0.052\ NO_x - 0.014\ Dpt + 0.888\ WD510 + 0.014\ WS510 + 0.456\ WD1019 + 0.009\ WS1019$	1993–1998	0.859

[a]The notations in the component are described in Table 1.

Database

Monitoring Site Location. The monitoring site was the WDNR monitoring site located at the University of Wisconsin, Milwaukee-North (UWM-N). This monitoring site was determined for the following reasons: (1) the Milwaukee area has been designated as a severe nonattainment area for the 1-hr National Ambient Air Quality Standards (NAAQS) for ozone since 1991, and (2) the air quality data at the UWM-N site are complete and of excellent quality compared with the other monitors.

The UWM-N monitoring station is at an elevation of 750 ft and is located at 43º 04.47′ N and 87º 53.07′ W. Several large multistoried buildings are on the north side of this monitoring station. A residential area surrounds the south side of the station. East of the station is a street with heavy traffic. The west side of the station is surrounded with either residential or urban areas.

Database Determination. Variables analyzed in this research include ozone, NO_x, and meteorological variables. The temporal domain is the 1991 ozone season. The ozone season officially is defined by the WDNR as April 15 through October 15, which takes into account all possible high-ozone days in a year. All of the ozone and NO_x observations were retrieved from the U.S. Environmental Protection Agency's (EPA) Aerometric Information Retrieval System (AIRS). Meteorological data were obtained through the State Climatology office and consisted of surface weather observations made by the National Weather Service (NWS) at the Milwaukee General Mitchell International Airport.

The 1991 ozone season is appropriate for this ozone study because

(1) 1991 is the calendar year having the greatest number of ozone 1-hr NAAQS exceedances for any year in Wisconsin during the 1990s. There were a relatively large number of exceedances observed in 1991 at UWM-N. From 1991 to 1999, 11 readings at UWM-N exceeded the 1-hr ozone standard (124 ppb when rounding is considered). However, six of them were observed in 1991. In 1991, 15 ozone readings were greater than 100 ppb and eight were greater than 120 ppb.

(2) During the summer of 1991, the Lake Michigan Ozone Study (LMOS) was conducted by the states of Wisconsin, Illinois, Michigan, and Indiana (see Figure 2). The LMOS field study yielded a full, rich, and accurate measurement database, which included data collected aboard ships in Lake Michigan, by aircraft, and at special ground-based monitoring sites. For 1991, considerable relevant information is available from the LMOS. Above all, the data quality of 1991 is reliable.

Predictor Variable Determination. The response variable and predictor variables are listed in Table 1. The variables used in this research were determined based on a comprehensive literature review.[7,8,36] Those candidate predictor variables generally include the four most important variables pointed out by Comrie:[15] daily maximum temperature, average dew-point temperature, average daily wind speed, and daily total sunshine. Daily values in this study were determined from hourly averages rather than from existing values that have been published. The 24-hr diurnal variations between ozone and the candidate predictor variables are displayed in Figure 4. June 26, 1991, was chosen to represent a typical high-ozone day.

The ozone-conducive weather conditions in the Lake Michigan Ozone Problem Area (see Figure 2) are usually associated with high temperature, high relative humidity, and southwesterly, southerly, or southeasterly winds along the Lake Michigan shore.[37] In most locations, peak daily temperature is highly correlated with ozone.[15,38,39] High dew-point temperatures are strongly associated with stagnating anticyclones and provide the moisture favorable for photochemical reactions.[15,38,39] Generally, low wind speeds and high solar radiation are significant contributors to high ozone.[15,38,39] In addition to the weather components, NO_x serves as an ozone precursor resulting from traffic emissions during the morning rush hour.[36]

In a synoptic meteorological situation, southeastern Wisconsin witnesses warm, humid, and weak southerly to southwesterly surface winds in the morning of high-ozone days. Fresh emissions from the morning rush hour and factories blow eastward over Lake Michigan. During the late spring and summer months, the surface waters of Lake Michigan are noticeably colder (by as much as 20 ºF) than the nearby land. Consequently, air parcels containing both regional ozone and ozone precursors sink toward the surface under clear skies.[40] Thus, as the subsiding air (rich in VOCs, NO_x, and increasing in ozone) descends to near Lake Michigan's surface, it slowly begins to spread horizontally toward the land in all directions. If these lake-based winds are strong enough to overcome the prevailing (land-based) synoptic-scale winds, they cause the onshore penetration of the ozone-laden "lake breeze" that usually occurs by late morning to mid-afternoon.

A back trajectory analysis indicated that almost 100% of all high-ozone days in Wisconsin occurred on days when the surface winds had a considerable southerly component.[3] Two general airflow patterns contribute to elevated ozone concentrations in eastern Wisconsin: a southerly-to-southeasterly path and a southwesterly direction, which help the ozone-laden lake breeze to occur. Consequently, wind variables are represented as two components to explain the lake breeze effect on ozone formation. The first wind component for the hourly periods of 5:00 a.m.–10:00 a.m. underscores precursor emission transport from the Milwaukee area to the lake in the mornings, and the second wind component for 10:00 a.m.–7:00 p.m. underscores ozone transport from the lake to the Milwaukee area in the late morning and afternoons.

Following Bloomfield et al.,[9] for each of the hourly periods, the two components of wind, v and u, are computed as the summation of the products of hourly wind speed times the cosine function of wind direction and as hourly wind speed times the sine function of wind direction, respectively. Then the v and u components are calculated by the vector average wind speed and wind direction specific to the 5:00 a.m.–10:00 a.m. and

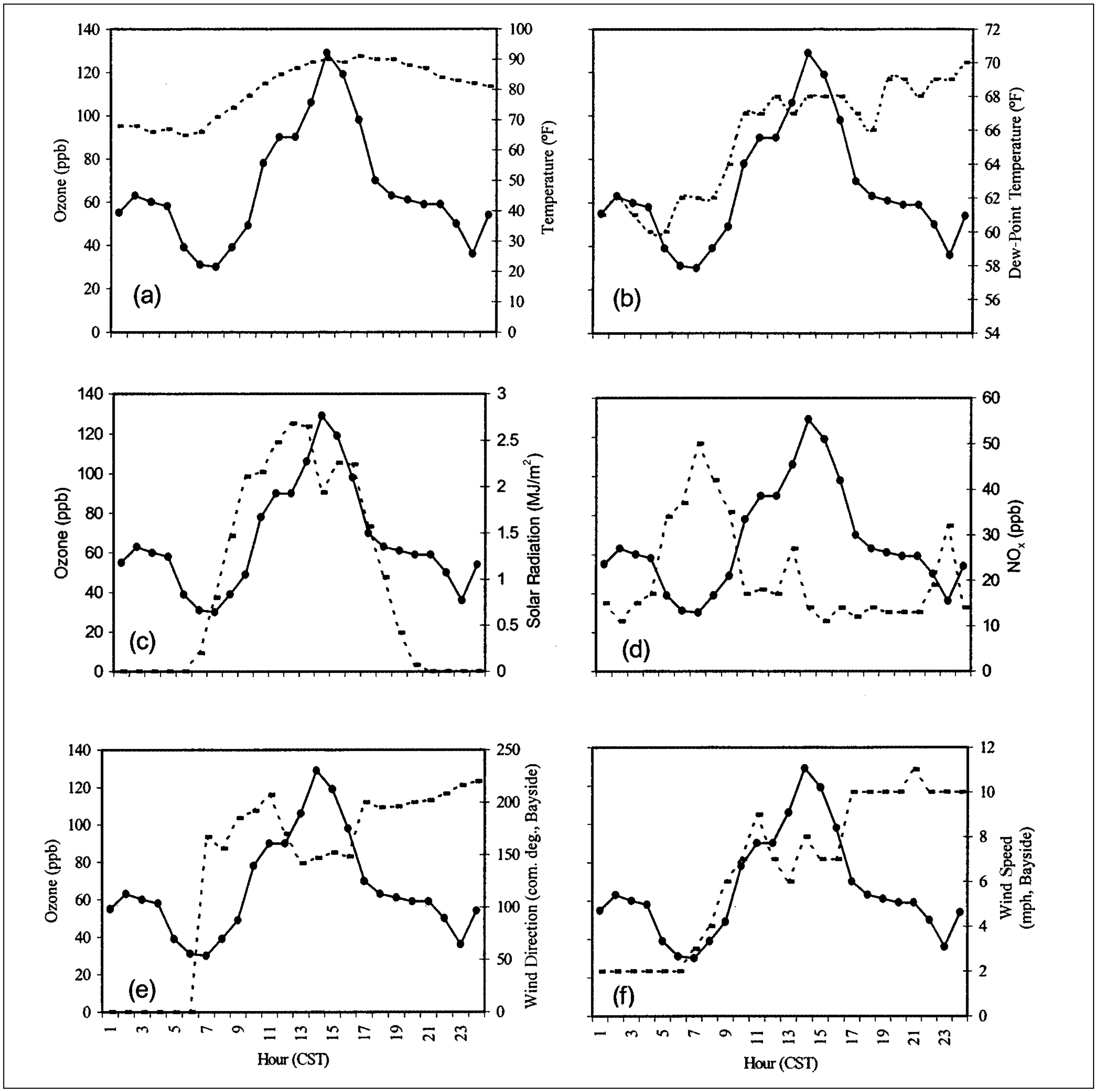

Figure 4. Hourly variation for ozone and the candidate predictor variables. Black dots with solid lines are ozone values. The dashed lines are (a) temperature, (b) dew-point temperature, (c) solar radiation, (d) NO_x, (e) wind direction, and (f) wind speed.

10:00 a.m.–7:00 p.m. time periods. One limitation on our predictor variables is that the hourly average of VOCs is not available. From a chemical standpoint, the production of ozone is a function of solar radiation and of the concentrations of the NO_x and VOCs emissions. Seinfeld and Pandis[41] emphasized that the ratio of NO_x/VOCs is also significant. Though NO_x and VOCs emissions are precursors of ozone pollution, several studies cited that emissions are not the dominant factor on ozone variation. Comrie and Yarnal[39] believed that much of the ozone variation is weather-related and emphasized the strong dependency of photochemical reactions on meteorological conditions. Also, Luria et al.[42] conclude that emissions from many major anthropogenic sources do not fluctuate significantly from day to day during weekdays. They highlighted that the dominant factors in ozone production are atmospheric conditions. Even though, for some studies, emissions were originally considered as possible predictor variables, they were finally discarded as not statistically significant.[23] Therefore, it is unlikely that the unavailability of emission data will prevent useful ozone predictions.

RESULTS

The RTSE Model with and without a PC Trigger

Temperature, WD510, and PC trigger were the resultant significant predictors out of the nine candidate predictors in the RTSE model (see Table 1). The idea of turning the PC trigger on and off is described in the flow chart in Figure 5. The PC trigger was employed in the RTSE model as one of the predictor variables, and its calculation is subject to whether the ozone level is greater than 100 ppb. The Pearson sample correlation between O_{pre} and O_{obs}, or its square $R^2 = r^2(O_{pre,} O_{obs})$, is used as a descriptive measure of the closeness of fit. Root Mean Square Error (RMSE) also is used as the performance statistic to evaluate the model.

Three time-series parameters show significant as well as temperature, WD510, current-day PC, and previous-day PC triggers in the RTSE model with a PC trigger.

$$\ln O3_t = 0.52 + 0.02Tp_t - 0.0003WD510_t + (0.001 + 0.0006B)PC_t + (1 - 0.20B^3)\, a_t/(1 - 0.47B - 0.28B^5) \quad (3)$$

The statistics of this model are addressed in Table 3. After predicting on the ln-scale, each forecast was transformed to the original scale by using eq 4, and r, R^2, and RMSE were calculated based on the original scale.

$$O_{pre} = \exp(O_{pre} + S.D. \times \frac{1}{2} S.D.) \quad (4)$$

where *S.D.*, the standard deviation of the residuals, equals 0.223.

The RMSE ($\hat{\sigma}_{a_t}$) is 12.2. The RTSE with a PC trigger model indicates a fairly good fit, with a correlation coefficient of 0.899 between fitted and observed ozone. The square correlation coefficient (R^2) is 0.809. The scatter plot of the observed and fitted ozone is displayed in Figure 6.

Table 3. Model statistics of the RTSE model with a PC trigger.

Parameters[a]	Estimate of Parameters	Standard Errors	T Ratio
Tp_t	0.02	0.002	11.41
$WD510_t$	–0.0003	0.0002	–1.94
PC_t	0.001	0.0002	6.02
PC_{t-1}	–0.0006	0.0002	–2.81
θ_3	0.20	0.080	2.55
ϕ_1	0.47	0.065	7.26
ϕ_5	0.28	0.065	4.35

[a] PC_{t-1} denotes the B back shift so that $BPC_t = PC_{t-1}$. Greek letters denote the parameters of the noise term N_t (eq 1).

Eq 3 is the form used for forecasting in a related paper.[2] The PC trigger will be calculated only when a high-ozone day is predicted. The RTSE model with a PC trigger shows a significant capability of predicting extremely high ozone. The improvement in estimating high ozone, which results from adding a PC trigger to the RTSE model, can be highlighted in comparison with an RTSE model without a PC trigger. Figures 7 and 8 show the significant difference between the two RTSE models with and without the PC trigger.

The RTSE model without PC resulted in only one significant variable, temperature, which is frequently recognized as the most important variable in ozone modeling. The resulting correlation coefficient is 0.825, the R^2 is

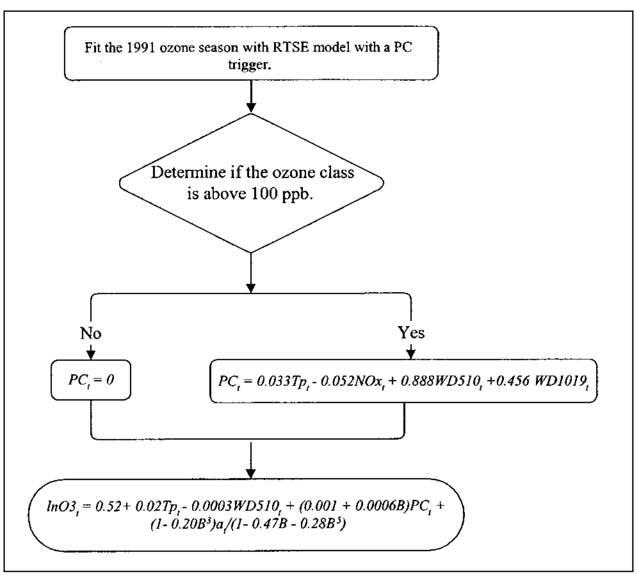

Figure 5. Process of fitting the 1991 ozone season with the RTSE model with a PC trigger.

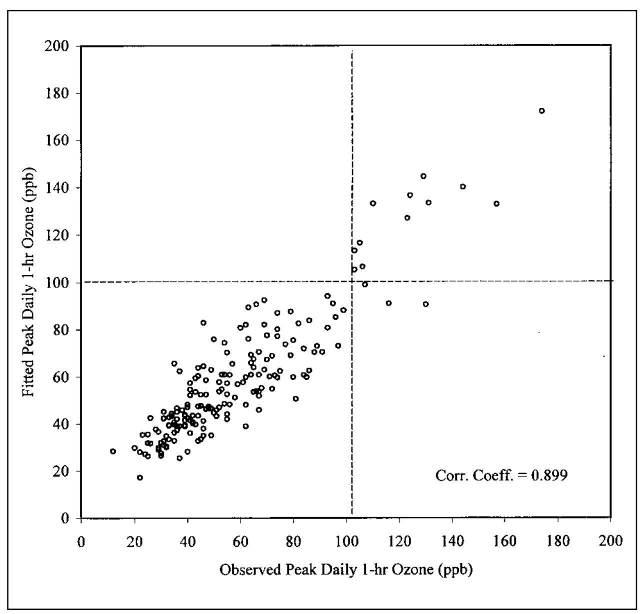

Figure 6. The RTSE model with the PC trigger, scatter plot of observed and fitted ozone.

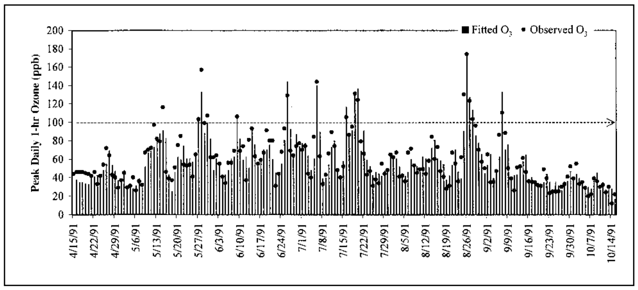

Figure 7. The RTSE model with the PC trigger fitting 1991 ozone series, Milwaukee.

0.680, and the RMSE is 15.6, on the original scale. The model is described in eq 5.

$$\ln O3_t = 0.70 + 0.03Tp_t + a_t/(1 - 0.55B + 0.16B^3 - 0.26B^5) \quad (5)$$

Figure 8 shows that prediction of ozone using eq 5 has the common difficulty of being unable to predict high ozone. This is particularly true for ozone greater than 100 ppb. The scatter plot of the observed and fitted ozone is shown in Figure 9. Table 4 lists the statistics of the RTSE model with and without PC for the entire ozone series. The number of predicted days with ozone above 100 ppb is apparently different between the two approaches. The RTSE model with PC predicted 12 out of 15 observed days with ozone above 100 ppb, but the RTSE model without PC predicted only five of the 15 days. Box and Jenkins[12] suggest using a 50% forecast interval instead of a 95% forecast interval when making forecasts with time-series models. The RTSE model with PC has a narrower 50% forecast interval (8.2 ppb) than that of the RTSE model without PC trigger (10.5 ppb).

Table 5 is a comparison between this research and the other relevant studies. Compared with the existing ozone models, R^2 of the RTSE model with PC seems to be relatively high, although the R^2 produced in this study was based on the same data that were used to build and validate the model. Chang and Cardelino[16] performed a

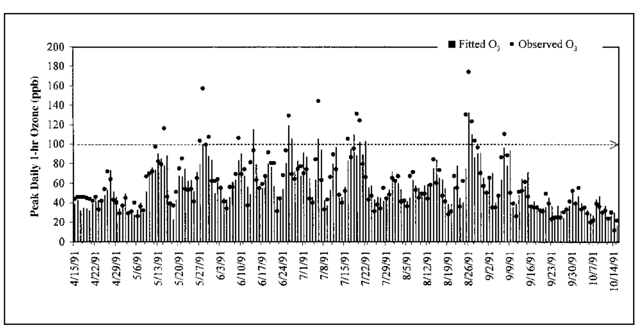

Figure 8. A pure RTSE model without PC trigger fitting 1991 ozone series, Milwaukee.

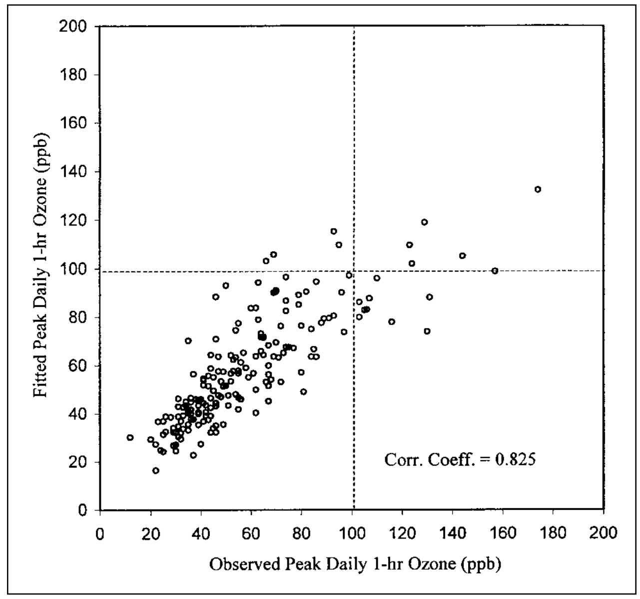

Figure 9. The RTSE model without the PC trigger, scatter plot of observed and fitted ozone.

real-time forecast for the summer of 1997. When forecasting real-time ozone, the result is limited greatly by the accuracy of weather forecasts that are major inputs to the model.[46] In other words, Chang and Cardelino[16] must have dealt with significant variability among the weather forecasts. Most of the remaining studies used different data sets to validate their models.

CONCLUSION

The proposed RTSE model with a PC trigger was applied to fit historical data. The results suggest that the problem of underpredicting high ozone has been reduced. To construct a useful ozone model, the following considerations should be addressed:

(1) Ambient air conditions, especially weather conditions that cause extremely high ozone, are different than those that cause normal ozone levels. Predicting extremely high ozone must be conducted differently than predicting normal ozone levels.

(2) Time-series correlation among ozone and other predictors cannot be ignored. A multivariate time-series model can describe persistence among ozone, ozone precursors, and meteorological factors. It also helps in predicting next-day ozone based on appropriate statistical assumptions.

(3) Statistical models have the advantage of easy operation, and they are relatively inexpensive and less time-consuming. Quantitative predictions can be generated with statistical models, and a forecast interval can be provided for each prediction.

Table 4. Statistics of RTSE model with and without PC trigger.

	Corr. Coeff.	R^2	RMSE	Predicted $O_3 \geq$ 100 ppb Days[a]	50% Forecast Interval
RTSE with PC	0.899	0.809	12.2	12	8.2
RTSE without PC	0.825	0.680	15.6	5	10.5

[a]The total number of days when ozone was greater than 100 ppb was 15.

Table 5. Ozone study review based on R^2 results.

Source	R^2	Method
Chang and Cardelino[16]	0.38	UAM-FM
Chen et al.[43]	0.45	Multidimensional phase space model
Comrie[15]	0.70	Neural network using unlagged variables
Davis and Speckman[44]	0.61–0.68	Loess/generalized additive model
Hubbard and Cobourn[8]	0.705–0.818	Hi-Lo + baseline hybrid regression model
Liu[45]	0.680	RTSE model without a PC trigger
Simpson and Layton[24]	0.608	Bivariate time-series model
This study	0.809	RTSE model with a PC trigger

The PC trigger described in eq 2 underlines the linear relationship among the meteorological and NO_x conditions for those days with ozone greater than 100 ppb. Using the PC trigger to improve high ozone predictions is an easy and useful approach. Adding the PC trigger to the RTSE model raised the R^2 from 0.680 to 0.809. For 15 days with ozone above 100 ppb, the RTSE model with PC trigger predicted 12 days, while the RTSE model without PC trigger only predicted five days.

The RTSE model with a PC trigger indicated its potential to predict normal and high ozone levels. The operation of this model to real-time forecasting will be detailed in a related paper.[2] Several PC trigger rules, which can determine the turning on or turning off the PC trigger, are developed in that paper.[2]

ACKNOWLEDGMENTS

The authors would like to extend our appreciation to the Wisconsin Department of Natural Resources for funding this publication, providing the monitoring data and constructive information, and funding this project. Pao-Wen G. Liu also thanks Professor Paul M. Berthouex, University of Wisconsin, Madison, for his advice and Larry Bruss, WDNR, for his input in this research.

REFERENCES

1. Draper, N.R.; Smith, H. *Applied Regression Analysis,* 2nd ed.; John Wiley & Sons: New York, 1981.
2. Liu, P.W.; Johnson, R. Forecast Peak Daily Ozone Levels—II. Using A Regression with Time Series Errors Model Having a Principal Component Trigger to Forecast 1999 Ozone Levels; *J. Air & Waste Manage. Assoc.,* accepted for publication.
3. *Attainment Demonstration for Ozone for the Year 2007, The Phase 3 Attainment State Implementation Plan (SIP) for the Eastern Wisconsin Nonattainment Areas*; Wisconsin Department of Natural Resources: December 2000.
4. Smith, B.E.; Adamski, W.J. Eight-Hour Ozone Trends of Sites in Lake Michigan Ozone Nonattainment Areas; *J. Air & Waste Manage. Assoc.* **1998**, *48*, 1204–1206.
5. Koletzke, M. M.S. Thesis, University of Central Florida, 1996.
6. Howard, C.M.; Barb, B.L. *Ozone Forecasting for Birmingham, Alabama;* Alabama Department of Environmental Management Air Division: Montgomery, AL, 1994.
7. Jirik, A.L. M.A. Thesis, University of Illinois, Chicago Circle, IL, 1978.
8. Hubbard, M.C.; Cobourn, W.G. Development of a Regression Model to Forecast Ground-Level Ozone Concentration in Louisville, KY; *Atmos. Environ.* **1998**, *32*, 2637–2647.
9. Bloomfield, P.; Royle, J.A.; Steinberg, L.J.; Yang, Q. Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends; *Atmos. Environ.* **1996**, *30*, 3067–3077.
10. Feister, U.; Balzer, K. Surface Ozone and Meteorological Predictors on a Subregional Scale; *Atmos. Environ.* **1991**, *25A*, 1781–1790.
11. Cox, W.M.; Chu, S. Meteorologically Adjusted Ozone Trends in Urban Areas: A Probabilistic Approach; *Atmos. Environ.* **1993**, *27B*, 425–434.
12. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis—Forecasting and Control,* 3rd ed.; Prentice-Hall: Englewood Cliffs, NJ, 1994.
13. Johnson, R.; Wichern, D.W. *Applied Multivariate Statistical Analysis;* Prentice Hall: Englewood Cliffs, NJ, 1992.
14. Cox, D.D.; Ensor, K.B. An Empirical Method for Prediction of Ambient Ozone Levels. In *Proceedings of the 88th Annual Meeting of the A&WMA*, San Antonio, TX, June 18–23, 1995.
15. Comrie, A.C. Comparing Neural Networks and Regression Models for Ozone Forecasting; *J. Air & Waste Manage. Assoc.* **1997**, *47*, 653–663.
16. Chang, M.E.; Cardelino, C. Application of the Urban Airshed Model to Forecasting Next-Day Peak Ozone Concentrations in Atlanta, Georgia; *J. Air & Waste Manage. Assoc.* **2000**, *50*, 2010–2024.
17. Lake Michigan Air Directors Consortium. *Midwest Subregional Modeling: 1-Hour Attainment Demonstration for Lake Michigan Area, Summary;* Illinois Environmental Protection Agency; Indiana Department of Environmental Management; Michigan Department of Environmental Quality; and Wisconsin Department of Natural Resources: September 18, 2000.
18. Adamski, W. *An Analysis of Measured and Predicted Concentrations Aloft of Ozone and NO_2 in the Lake Michigan Airshed during 12–14 July, 1995*; Bureau of Air Management, Wisconsin Department of Natural Resources: 1997.
19. Rao, S.T.; Zurbenko, I.G. Detecting and Tracking Changes in Ozone Air Quality; *J. Air & Waste Manage. Assoc.* **1994**, *44*, 1089–1092.
20. Rao, S.T.; Zurbenko, I.G.; Porter, P.S.; Ku, J.Y.; Henry, R.F. Dealing with the Ozone Nonattainment Problem in the Eastern United States; *EM* **1996**, *January*, 17–31.
21. Cobourn, W.G.; Hubbard, M.C. Nonlinear Regression and Trajectory Analysis Applied to Ozone Forecasting in Louisville, Kentucky. In *Proceedings of the 92nd Annual Meeting of the A&WMA*, St. Louis, MO, June 20–24, 1999.
22. Jorquera, H.; Perez, R.; Cipriano, A.; Espejo, A.; Letelier, M.V.; Acuna, G. Forecasting Ozone Daily Maximum Levels at Santiago, Chile; *Atmos. Environ.* **1998**, *32*, 3415–3424.
23. Robeson, S.M.; Steyn, D.G. Evaluation and Comparison of Statistical Forecast Models for Daily Maximum Ozone Concentrations; *Atmos. Environ.* **1990**, *24B*, 303–312.
24. Simpson, R.W.; Layton, A.P. Forecasting Peak Ozone Levels; *Atmos. Environ.* **1983**, *17*, 1649–1654.
25. Prior, E.J.; Schiess, J.R.; McDougal, D.S. Approach to Forecasting Daily Maximum Ozone Levels in St. Louis; *Environ. Sci. Technol.* **1981**, *15*, 430–436.
26. Milionis, A.E.; Davies, T.D. Regression and Stochastic Models for Air Pollution—I. Review, Comments and Suggestions; *Atmos. Environ.* **1994**, *28*, 2801–2810.
27. Milionis, A.E.; Davies, T.D. Regression and Stochastic Models for Air Pollution—II. Application of Stochastic Models to Examine the Links between Ground-Level Smoke Concentrations and Temperature Inversions; *Atmos. Environ.* **1994**, *28*, 2811–2822.

28. Flaum, J.B.; Rao, S.T.; Zurbenko, I.G. Moderating the Influence of Meteorological Conditions on Ambient Ozone Concentrations; *J. Air & Waste Manage. Assoc.* **1996**, *46*, 35–46.
29. Rao, S.T.; Zalewsky, E.; Zurbenko, I.G. Determining Temporal and Spatial Variations in Ozone Air Quality; *J. Air & Waste Manage. Assoc.* **1995**, *45*, 57–61.
30. Cobourn, G.W.; Dolcine, L.; French, M.; Hubbard, M. A Comparison of Nonlinear Regression and Neural Network Models for Ground-Level Ozone Forecasting; *J. Air & Waste Manage. Assoc.* **2000**, *50*, 1999–2009.
31. *SAS/ETS Software: Applications Guide 1 Version 6: Time Series Modeling and Forecasting, Financial Reporting, and Loan Analysis,* 1st ed.; SAS Institute Inc.: Cary, NC, 1991.
32. Pankratz, A. *Forecasting with Univariate Box-Jenkins Models—Concepts and Cases*; John Wiley & Sons: New York, 1983.
33. Liu, L.; Hudak, G.B.; Box, G.E.P.; Mervin, E.M.; Tiao, G.C. *Forecasting and Time Series Analysis Using the SCA Statistical System*; Scientific Computing Associates: Dekalb, IL, 1992.
34. *SAS/STAT User's Guide, Version 6,* 4th ed.; SAS Institute Inc.: Cary, NC, 1990; Vol. 1.
35. Ryan, W.F. Forecasting Severe Ozone Episodes in the Baltimore Metropolitan Area; *Atmos. Environ.* **1995**, *29*, 2387–2398.
36. Yi, J.; Prybutok, V.R. A Neural Network Model Forecasting for Prediction of Daily Maximum Ozone Concentration in an Industrial Urban Area; *Environ. Pollut.* **1996**, *92*, 349–357.
37. *Operation Ozone Cleaning the Air in Wisconsin;* Wisconsin Department of Natural Resources: 2000; Vol. III, No. 2.
38. Comrie, A.C. The Climatology of Surface Ozone in Rural Areas: A Conceptual Model; *Prog. Phys. Geog.* **1990**, *14*, 295–316.
39. Comrie, A.C.; Yarnal, B. Relationships between Synoptic-Scale Atmospheric Circulation and Ozone Concentrations in Metropolitan Pittsburgh, Pennsylvania; *Atmos. Environ.* **1992**, *26B*, 301–312.
40. Keen, C.S.; Lyons, W.A. Lake/Land Breeze Circulation on the Western Shore of Lake Michigan; *J. Appl. Meteor.* **1978**, *17*, 1843–1855.
41. Seinfield, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics—From Air Pollution to Climate Change*; John Wiley and Sons: New York, 1998.
42. Luria, M.; Boatman, J.F.; Wellman, D.L.; Gunter, R.L.; Watkins, B.A.; Wilkion, S.W.; Van Valin, C.C. Lake Michigan Ozone Study (LMOS): Measurements from an Instrumental Aircraft; *Atmos. Environ.* **1992**, *26A*, 3265–3277.
43. Chen, J.; Islam, S.; Biswas, P. Nonlinear Dynamics of Hourly Ozone Concentrations: Nonparametric Short-Term Prediction; *Atmos. Environ.* **1998**, *32* (11), 1839–1848.
44. Davis, J.M.; Speckman, P. A Model for Predicting Maximum and 8-hr Average Ozone in Houston; *Atmos. Environ.* **1999**, *33* (16), 2487–2500.
45. Liu, P.W. Designing a Method for Forecasting Daily Maximum Ozone Concentrations in Milwaukee. A report submitted in partial fulfillment of the requirements for the Ph.D. preliminary examination; University of Wisconsin, Madison, WI, December 1998.
46. Liu, P.W. Ph.D. Dissertation, University of Wisconsin, Madison, WI, May 2000.

APPENDIX—TIME SERIES AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODEL

To illustrate typical classes of time-series models cited in this paper, AR(p) and ARMA(p,q) models are explained as follows.

Autoregressive Process of Order p [AR(p)]

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + a_t \qquad (A.1)$$

where a_t is white noise and $\phi_1,...,\phi_p$ are parameters which must lie within certain limits. Here p stands for the lag that counts down to the day of $t - p$. Equation A.1 can be written as

$$\left(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p\right) X_t = a_t \qquad (A.2)$$

where B is the backshift operator [i.e., $B^k X_t = X_{t-k,}$ and $(1-B^k)X_t = X_t - X_{t-k}$ for $k = 1, .., p$].

Mixed Autoregressive Moving Average Processes of Order p, q [ARMA(p, q)]

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} \\ &+ a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \end{aligned} \qquad (A.3)$$

It may be written as

$$\begin{aligned} &(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) X_t \\ &= (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q) a_t \end{aligned} \qquad (A.4)$$

In an ARMA(p, q) model, the autoregressive (AR) term has the order p and the moving average (MA) term has the order q. The AR term is on the top line of eq A.4 and the MA term is on the bottom line of eq A.4.

About the Authors

Pao-Wen Grace Liu is an air pollution control specialist at the Bureau of Air Management, Wisconsin Department of Natural Resources. She received her Ph.D. from the University of Wisconsin, Madison, by completing this ozone-forecasting project. Richard Johnson is a professor of statistics at the University of Wisconsin, Madison. The corresponding author is Pao-Wen Grace Liu, e-mail: Pao-Wen.Grace@dnr.state.wi.us.