Atmospheric Environment Vol. 23, No. 3, pp. 689–692, 1989.
Printed in Great Britain.

0004–6981/89 $3.00+0.00
© 1989 Pergamon Press plc

A CONDITIONAL PROBABILITY DENSITY FUNCTION FOR FORECASTING OZONE AIR QUALITY DATA

S. M. ROBESON* and D. G. STEYN

Department of Geography, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5

(*First received* 27 *October* 1987 *and in final form* 15 *August* 1988)

Abstract—Probabilistic forecasts are often employed to estimate the potential for high pollutant concentrations. To develop a probabilistic forecast of ozone concentrations, we suggest that use be made of the inherent properties of seasonality and autocorrelation in O_3 time series. A non-stationary, autocorrelated stochastic process is used to simulate a conditional probability density function (p.d.f.) which quantifies the effects of seasonality and autocorrelation. To illustrate the utility of such a model, the simulated conditional p.d.f. is shown to be clearly superior to an ordinary p.d.f. developed from summer ozone data.

Key word index: O_3, autocorrelation, seasonality, simulation, non-stationarity, forecasting.

1. INTRODUCTION

In order to avoid potentially hazardous tropospheric O_3 levels (particularly near densely populated areas), accurate forecasts of the atmospheric concentration of O_3 are needed. In most cases, models which forecast specific point values of pollutants are employed (e.g. McCollister and Wilson, 1975; Wolff and Lioy, 1978; Aron, 1980; Prior *et al.*, 1981; Simpson and Layton, 1983). Often, however, a probabilistic forecast (i.e. the probability of exceeding a given concentration level) is much more easily interpreted and therefore provides greater utility.

A probability density function (p.d.f.) may be fitted to previously observed values in order to determine the probability of exceeding air quality standards (see Bencala and Seinfeld, 1976). While this approach is acceptable for long-term emission control strategy, the use of an ordinary p.d.f. assumes that observed values are (1) independent of one another and (2) derived from a stationary time series. Ozone concentrations are rarely independently distributed or stationary—several studies of the annual variability of O_3 in the urban troposphere have shown both strong seasonal dependence and serial correlation (see Merz *et al.*, 1972; Chock *et al.*, 1975; Horowitz and Barakat, 1979; Hirtzel and Quon, 1981; Robeson, 1987). Although Horowitz and Barakat (1979) have theoretically shown that the presence of autocorrelation in sequences of pollutant data does not affect the computation of an annual p.d.f. for daily maximum concentrations, a conditional p.d.f. (i.e. the p.d.f. for tomorrow's O_3 value given past conditions) is certainly affected by both autocorrelation and seasonality. In other words, high or low concentrations tend to occur in sequence and at certain times during the year. Hence, the use of an ordinary p.d.f. for probabilistic forecasts of O_3 concentrations is highly inappropriate. We suggest that use be made of the inherent seasonality and serial correlation of O_3 data by employing a non-stationary, stochastic model to generate a conditional p.d.f. of the daily maximum 1-h average concentration. By simulating the inherent properties of O_3 time series, a dynamic p.d.f. is developed.

2. METHOD

The conditional p.d.f. to be developed here allows one to determine the probability of exceeding a given O_3 concentration level by accounting for both seasonality and serial correlation in the O_3 time series. Alternative versions of conditional p.d.f.'s may also be utilized. The conditional p.d.f. of Hirtzel and Quon (1981) gives the probability of "an exceedance continuing for t_2 days given the exceedance has lasted t_1 days". Hence, their model fixes the O_3 concentration and determines the probability of a specified exceedance duration while our model fixes the time element (i.e. the model forecasts tomorrow's p.d.f.) and determines the probability of exceeding a given concentration.

Assuming that O_3 is generated by a non-stationary, autocorrelated stochastic process (pollutant concentrations are usually assumed to be generated by a random process), the model described by Horowitz and Barakat (1979) may be used to simulate O_3 variability. Horowitz and Barakat (1979) examined one year of O_3 data from the St. Louis Regional Air Pollution Study to show the utility of such a model (for other purposes). Although a single year of data

*Present affiliation: Center for Climatic Research, Department of Geography, University of Delaware, Newark, DE 19716, U.S.A.

exhibits a large degree of non-stationarity due to the seasonal variability of O_3, year-to-year variation due to synoptic-scale weather conditions and/or emissions trends may induce other types of non-stationarity; therefore, depending upon the data used, other methods than those described by Horowitz and Barakat (1979) may be needed to render a time series stationary.

In the present study, the 3-parameter log-normal procedure outlined by Ott and Mage (1976) is first used to transform the original data to a Gaussian distribution. To reproduce the seasonal trend of daily O_3 maxima, a polynomial is fitted via ordinary least-squares estimation. The use of a polynomial to describe seasonal variability is justified since the least-squares procedure objectively determines the timing of the seasonal maximum. Discontinuities in the polynomial may occur in the transition from 31 December (day number 365 or 366) to 1 January (day number 1); however, there should be little impact since O_3 concentrations are generally quite low in winter. (For locations in the Southern Hemisphere, the day numbering scheme should be altered.) After removal of the seasonal trend, a strongly autocorrelated sequence remains. This sequence may be described by

$$e_t = \ln(X_t - K) - \left[\sum_{i=0}^{n} b_i t^i\right], \qquad (1)$$

where e_t is the residual series, X_t is the daily O_3 maximum on day t, K is a constant, and the b_i are coefficients from a polynomial regression of order n. The time series method of Box and Jenkins (1976) is used to filter the e_t series into a Normally and Independently Distributed (NID) sequence, a_t:

$$\begin{aligned} w_t = \phi_1 w_{t-1} + \ldots + \phi_p w_{t-p} + a_t \\ -\theta_1 a_{t-1} - \ldots - \theta_{q^{a_{t-q}}} + \theta_0 \end{aligned} \qquad (2)$$

where $\phi_1, \ldots, \phi_p$ are the autoregressive terms, $\theta_1, \ldots, \theta_q$ are the moving average terms, θ_0 is a constant, and w_t is the differenced original series,

$$w_t = \nabla^s e_t. \qquad (3)$$

The backward difference operator, ∇, is defined as

$$\nabla^s e_t = e_t - e_{t-s}, \qquad (4)$$

which may be repeated d times. Using the terminology of Box and Jenkins (1976), an Auto-Regressive Integrated Moving Average model with p autoregressive terms, d differencing operations, and q moving average terms is designated ARIMA (p, d, q).

The a_t terms are "random shocks" or residuals (with mean value of zero) which incorporate the effects of all the factors other than past time series values which act to influence O_3 concentrations. If model identification is performed properly, the a_t sequence is NID and is therefore easily simulated via a NID random number generator. Once parameters have been estimated, this extremely compact representation (i.e. the combination of Equations 1–4) of system variability is able to simulate a conditional p.d.f. for tomorrow's daily maximum O_3 concentration when two pieces of information are known: time of year and today's daily maximum O_3 concentration. The number of simulated realizations of tomorrow's daily maximum O_3 concentration is subjectively determined by choosing the length of the a_t series. Since Equations 1–4 are not very computationally demanding, a large sample is preferable.

It is interesting to note that Chock (1986) has used a similar method to that outlined above for the purposes of simulating extreme values of air quality data. Although he found that such a model does not simulate extreme values with great accuracy or confidence, for less extreme values, the model performs well. He concluded that less extreme values should be used for air quality criteria, but it may also be said that the model he described may be valuable for probabilistic forecasting of all but the most extreme values.

3. RESULTS

To illustrate the utility of the model for a conditional p.d.f. described above, O_3 data from the Greater Vancouver region of British Columbia, Canada (population of approximately 1.5 million) is analyzed. During summer, daily O_3 maxima in this area have commonly exceeded the "maximum acceptable" level (82 ppb) and have occasionally exceeded the "maximum tolerable" level (153 ppb) as established by the Canadian National Air Quality Objectives (CSC, 1982, 1985). A monitoring station (designated station T9 by the Greater Vancouver Regional District, Pollution Control Section) located in Rocky Point Park alongside Burrard Inlet is a National Air Pollution Survey Class 1 station. Hourly averaged measurements have been made since 1978 using either chemiluminescence from the reaction with ethylene (Bendix Model 8002) or ultraviolet photometry (TECO Model 49). Robeson (1987) may be consulted for further description of monitoring practices at station T9.

Data from the period 1978–1985 were used to estimate system parameters. A second-order polynomial was sufficient to remove the seasonal trend while inspection of autocorrelation and partial autocorrelation functions clearly suggested an ARIMA (1, 0, 0) model (for all years) to describe the residual series e_t. Parameter estimates (with standard errors in parentheses) are $b_0 = 3.57$ (0.02), $b_1 = 8.58 \times 10^{-2}$ (2×10^{-4}), $b_2 = -2.45 \times 10^{-4}$ (6×10^{-7}), $\phi_1 = 0.526$ (0.02), $\theta_0 \approx 0$, $K = -25$, and a_t is NID with a mean value of approximately 0.0 and a standard deviation of 0.238.

A simulated realization of a conditional p.d.f. for a midsummer day (21 July) when the previous day's daily maximum O_3 concentration is 82 ppb is presented in Fig. 1. Given a finite data sample which has very few midsummer days with daily O_3 maxima of

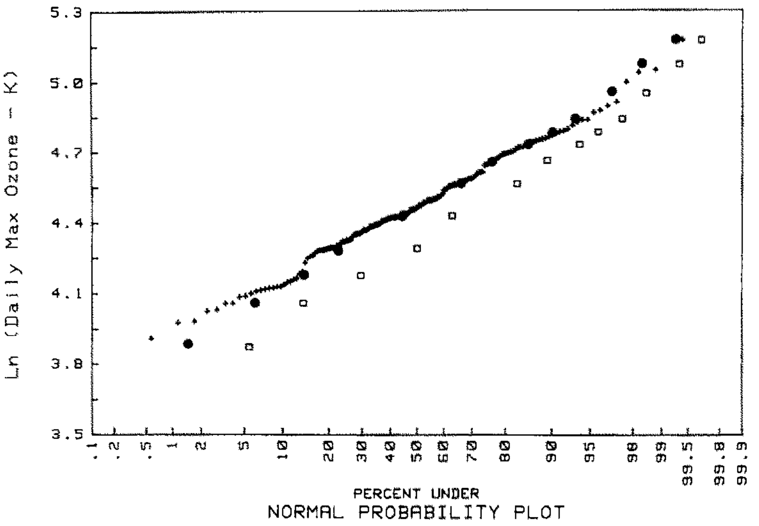

Fig. 1. Simulated and observed cumulative probabilities for station T9. Symbols used represent the following: +—simulation using Equations 1–4; ●—values observed at station T9 under conditions similar to those used in simulation (see text); □—all May to September data, 1978–1985.

82 ppb, the observed conditional p.d.f. shown in Fig. 1 is drawn from not only days with concentrations of 82 ppb, but from a range of concentrations (in this case, arbitrarily chosen to be from 64 to 100 ppb) on days in June–August, 1978–1985. Hence, when today's daily O_3 maximum is 82 ± 18, tomorrow's daily O_3 maximum is included in the observed conditional p.d.f. (The total number of values used in the observed conditional p.d.f. was 173.) The simulation uses today's value to compute the e_{t-1} term while the a_t series is formed using a normally distributed random series. The simulated conditional p.d.f. appears to match the observed values reasonably well. For comparison, the cumulative frequency distribution from values observed during high O_3 months (May–September) is also shown in Fig. 1. Clearly, the simulation which includes both seasonality and an autoregressive term is an improvement over simply compiling historical data to develop a p.d.f. for any given season.

4. CONCLUSIONS

Ozone concentrations are neither independently distributed nor stationary through time. Hence, in order to generate probabilistic forecasts of O_3 concentrations, it is proposed that use be made of properties derived from historical O_3 measurements at given sites. The model described by Horowitz and Barakat (1979) is used to generate a conditional p.d.f. which utilizes pertinent information regarding the O_3 time series: time of year (via a seasonal polynomial) and the previous day's daily O_3 maximum (via an autoregressive term). Incorporating information relevant to system variability into a conditional p.d.f. has been shown to be an improvement over disregarding inherent properties of O_3 time series.

Acknowledgements—The Pollution Control Section of the Greater Vancouver Regional District kindly provided the ozone data. The research was supported by a grant from the Natural Science and Engineering Research Council of Canada and by a University of British Columbia Graduate Fellowship to SMR.

REFERENCES

Aron R. (1980) Forecasting high level oxidant concentrations in the Los Angeles basin. *JAPCA* **30**(11), 1227–1228.

Bencala K. E. and Seinfeld J. H. (1976) On frequency distributions of air pollution concentrations. *Atmospheric Environment* **10**, 941–950.

Box G. E. P. and Jenkins G. M. (1976) *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Chock D. P. (1986) Statistics of extreme values of air quality—a simulation study. *Atmospheric Environment* **19**, 1713–1724.

Chock D. P., Terrell T. R. and Levitt S. B. (1975) Time-series analysis of Riverside, California air quality data. *Atmospheric Environment* **9**, 978–989.

Concord Scientific Corporation (CSC) (1982) *Vancouver Oxidant Study, Air Quality Analysis* 1978–1981. Prepared for Environment Canada, Environmental Protection Service.

Concord Scientific Corporation (CSC) (1985) *Vancouver Oxidant Study, Air Quality Analysis Update* 1982–1984. Prepared for Environment Canada, Environmental Protection Service.

Hirtzel C. S. and Quon J. E. (1981) Statistical analysis of continuous ozone measurements. *Atmospheric Environment* **15**, 1025–1034.

Horowitz J. and Barakat S. (1979) Statistical analysis of the maximum concentration of an air pollutant: effects of autocorrelation and non-stationarity. *Atmospheric Environment* **13**, 811–818.

Merz P. H., Painter L. J. and Ryason P. R. (1972) Aerometric data analysis—time series analysis and forecast and an atmospheric smog diagram. *Atmospheric Environment* **6**, 319–342.

McCollister G. M. and Wilson K. R. (1975) Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. *Atmospheric Environment* **9**, 417–423.

Ott W. R. and Mage D. T. (1976) A general purpose univariate probability model for environmental data analysis. *Computers and Operations Research* **3**, 209–216.

Prior E. J., Schiess J. R. and McDougal D. S. (1981) Approach to forecasting daily maximum ozone levels in St. Louis. *Envir. Sci. Technol.* **15**, 430–436.

Robeson S. M. (1987) *Time Series Analysis of Surface Layer Ozone in the Lower Fraser Valley of British Columbia.* M.Sc. Thesis, The University of British Columbia, Vancouver, British Columbia, Canada.

Simpson R. W. and Layton A. P. (1983) Forecasting peak ozone levels. *Atmospheric Environment* **17**, 1649–1654.

Wolff G. T. and Lioy P. J. (1978) An empirical model for forecasting maximum daily ozone levels in the northeastern U.S. *JAPCA* **28**, 1034–1038.