AIR QUALITY MODELS FOR ENVIRONMENTAL MANAGEMENT

by

John Ashley Taylor

A thesis submitted to the
Australian National University
for the degree of Doctor of Philosophy
December 1985

THE AUSTRALIAN
LIBRARY
NATIONAL UNIVERSITY

PREFACE

The following publications were prepared as the result of research undertaken jointly with Dr A.J. Jakeman and Dr R.W. Simpson.

Jakeman A.J., Taylor J.A. and Simpson R.W. (1984) An air quality modeling approach for environmental impact assessment, airshed management and pollution control policy. Proceedings of the Simulation Society of Australia Conference, University of Adelaide, August 13-15, 65-72.

Jakeman A.J., Simpson R.W. and Taylor J.A. (1984) A simulation approach to assess air pollution from road transport. IEEE Transactions on Systems, Man and Cybernetics 14, 726-736.

Taylor J.A., Simpson R.W. and Jakeman A.J. (1984) Predicting the spatial variation of maximum air pollutant concentrations from limited observations. CRES Working Paper 1984/19, Australian National University.

Jakeman A.J., Simpson R.W. and Taylor J.A. (1985) Combining deterministic and statistical models for ill-defined systems: Advantages for air quality asessment. Mathematics and Computers in Simulation 27, 167-178.

Taylor J.A. and Jakeman A.J. (1985) Identification of a distributional model. Communications in Statistics B14, 497-508.

Taylor J.A., Simpson R.W. and Jakeman A.J. (1985) A hybrid model for predicting the distribution of pollutants dispersed from line sources. Science of the Total Environment 46, 191-213.

Jakeman A.J. and Taylor J.A. (1985) A hybrid ATDL-gamma distribution model for predicting area source acid gas concentrations. Atmospheric Environment 19, 1959-1967.

Taylor J.A., Jakeman A.J. and Simpson R.W. (1985) Modelling distributions of air pollutant concentrations Part I - Identification of statistical models. Atmospheric Environment (in press).

Jakeman A.J., Taylor J.A. and Simpson R.W. (1985) Modelling distributions of air pollutant concentrations Part II - Estimation of the parameters of the lognormal, gamma, Weibull and exponential distributions. Atmospheric Environment (in press).

Taylor J.A., Simpson R.W. and Jakeman A.J. (1985) A hybrid model for predicting the distribution of sulphur dioxide concentrations observed near elevated point sources. (submitted to Ecological Modelling).

Taylor J.A., Simpson R.W. and Jakeman A.J. (1985) Statistical modeling of restricted pollutant data sets to assess compliance with air quality criteria. (submitted to Environmental Monitoring and Assessment).

The text of these papers has at times been closely followed in Chapters 3 to 8 of this thesis. The remainder of this thesis, except where otherwise acknowledged in the text, represents the original research of the author.

John Ashley Taylor
December 1985

ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr Tony Jakeman and Dr Rod Simpson for their encouragement and the interest they have shown in this work.

I am also grateful to the following individuals and organizations: to the Newcastle City Council Health Division who provided the acid gas data examined in Chapters 5 and 8; to the Victorian Environment Protection Authority and Angela Marles of the Department of Arts, Heritage and the Environment who provided air quality data recorded in Melbourne, the analysis of which is reported in Chapter 3; the Country Roads Board of Victoria, in particular the efforts of Sam Maccarone who provided the air quality data analysed in Chapter 6 and to Monsieur M. Benarie and Dr D.P. Chock who commented on an earlier draft of Chapter 6; and I would like to thank the Department of Conservation and Environment, particularly John Rosher, and the Department of Health, Western Australia, who provided the air quality data employed in the study as reported in Chapter 7.

I am also grateful to Janine Corey for her fast and accurate typing and to Sarah Titchen whose careful proofreading of the final draft ensured that the mistakes remaining are all my own.

Finally I would like to thank my family and friends for their constant encouragement and support throughout this course of study and to the staff of the Centre for Resource and Environmental Studies who provided an interesting and stimulating environment within which this research was undertaken.

This work would not have been possible without the assistance of a Commonwealth Postgraduate Research Award.

ABSTRACT

In this thesis mathematical models of air pollution concentrations are devised for the purpose of air quality management. The models constructed predict the entire distribution of concentration, although emphasis is given to the prediction of the upper percentiles as it is these concentrations which are most often referred to by air quality criteria. Models are developed which combine two key approaches to air quality modelling, namely deterministic and statistical modelling. They are linked in such a manner that the strengths of each approach are exploited and the weaknesses attenuated. This approach to air quality modelling is referred to here as the hybrid modelling approach. Statistical models have also been developed to assist in the formulation of monitoring programs which assess compliance with air quality standards ~~criteria~~ based upon complete and restricted data sets. All models in this thesis incorporate a level of complexity which is compatible both with the available data and with the objectives of the modelling exercise.

The methods of estimation of the parameters of the distributional model component of the hybrid models are considered and it is demonstrated how simple empirical models can be constructed which approximate the minimum levels of uncertainty associated with model predictions. The problem of identification of a distributional model for air quality data is also examined and it is shown that model identification performed using the maximum of the log likelihood functions in combination with modified Kolmogorov statistics selects the best distributional model with high probability from amongst the lognormal, gamma, Weibull and the exponential models. The model identification procedure is applied to a data set consisting of measurements of six pollutants recorded in a large urban area at a number of sites and over several years.

With these techniques of model identification and parameter estimation hybrid models were developed for three emission source regimes - area, line and point sources. The data sets considered include acid gas concentrations recorded in Newcastle, Australia, carbon monoxide levels observed near a roadway in Melbourne and sulphur dioxide levels produced by gold roasting and nickel smelting in Kalgoorlie. The hybrid models employ the variance-covariance matrix of the maximum likelihood parameter estimates in conjunction with Monte Carlo experimentation to generate approximate confidence intervals for model predictions. These confidence intervals were found to provide reasonable bounds upon model uncertainty.

The problem of assessing compliance with air quality standards where the data set may be both complete or incomplete is also addressed. The procedures of statistical model identification and parameter estimation were applied. Additionally a nonparametric procedure based upon an empirical quantile-quantile comparison of data at two monitoring sites was developed. The importance of model identification was clearly demonstrated. However, the empirical quantile-quantile model yielded the best results and should be employed where applicable.

TABLE OF CONTENTS

Chapter 7: A HYBRID MODEL FOR PREDICTING THE DISTRIBUTION OF SULPHUR DIOXIDE CONCENTRATIONS OBSERVED NEAR ELEVATED POINT SOURCES

Chapter 8: ASSESSING COMPLIANCE WITH AIR QUALITY CRITERIA USING STATISTICAL MODELS OF RESTRICTED DATA SETS

CHAPTER 1
AIR QUALITY - THE STUDY OF A COMPLEX SYSTEM

1.1 Introduction

The introduction in the United States of the Clean Air Act Amendments of 1970 and 1977, which included the prevention of significant deterioration, air quality maintenance plans and new source permits, initiated a substantial and continuing research effort into the development of air quality models in that country. These amendments saw the adoption of an approach to the control of air pollution which defined air quality management as the regulation of pollutant emissions in such a manner as to achieve a specified set of national ambient air quality standards or goals (De Nevers et al., 1977). This definition implies, along with the statement of air quality goals, that estimates of the pollutant emissions, observations of ambient air quality and models for the dispersion of air pollutants, be available. This air quality management approach to the control of air pollution has been adopted in numerous countries (Campbell and Heath, 1977).

Mathematical models for the dispersion of air pollutants may be constructed for scientific understanding with an aim to explain the complex detail of the physical and chemical processes involved. Alternatively models may be developed for the purpose of air quality management. While the former approach can involve considerable complexity of description, the latter requires only that detail comensurate with the aims of air quality management and sympathetic to the available a priori information be incorporated within the model.

The aim of this thesis is to demonstrate that simple but effective models can be constructed for the purpose of air quality management in spite of problems associated with poor data and in some cases limited knowledge. The models are used to predict ambient pollutant concentrations in a form that allows direct comparison with air quality criteria. The models are simple only in the sense that they contain as few parameters as are necessary for good prediction. Importantly the models developed yield estimates of the uncertainty associated with model predictions. This form of model output, while suited to other applications, is of particular importance to the problem of air quality management where accurate predictions are not generally possible.

1.2 The air pollution problem

The term 'air pollution' is defined here to mean the addition of any substance to the atmosphere from sources which directly or through transformation is present at a concentration sufficiently above normal ambient levels to produce a measurable effect upon humans, ecosystem or materials. A pollutant may be a substance that has been manufactured or is naturally occurring and may be a gas, aerosol or solid.

Gaseous pollutants added directly to the atmosphere include the oxides of nitrogen, sulphur and carbon to name but a few. These pollutants are usually termed primary pollutants whereas the secondary pollutants, including ozone and compounds derived from the sulphur and nitrogen oxides, are produced in the atmosphere by chemical reaction. These chemical reactions can take place between the primary pollutants, the normal constituents of the atmosphere and other secondary pollutants. Secondary pollutants are also produced by the decay of radioactive substances.

Perhaps the most well known form of secondary pollution is collectively known as photochemical smog. Unlike the primary pollutants such as sulphur dioxide and particulates which are produced by the combustion of coal, photochemical smog represents a complex series of reactions involving nitrogen oxides, non-methane hydrocarbons and sunlight (Seinfeld, 1975). Photochemical smog is generally associated with cities possessing large motor vehicle populations, poor dispersive conditions such as occur in river valleys and high solar radiation. Los Angeles is probably the best known of the cities experiencing photochemical smog and was the first to report a severe episode for which acute eye iritation and sore throats were indicators. In Australia the cities of Sydney, Melbourne and Brisbane experience high levels of photochemical smog (Lawlor, 1982).

While it would be desirable to eliminate air pollution altogether, economic constraints and the presence of natural background pollutant concentrations makes this task all but impossible. Australia currently contributes to a world-wide network of monitoring stations whose purpose is to monitor the global background concentrations of pollutants including carbon dioxide, carbon monoxide, methane, ozone, nitrogen oxides

Table 1.1: Natural source and background concentrations for selected pollutants for which the World Health Organization has established ambient air quality standards.

Pollutant	Natural Source	Background concentration (μgm^{-3})
Sulphur dioxide	Volcanoes	1-4
Nitrogen oxide	Bacterial action in soil; photodissociation of N_2O and NO_2	0.3-2.5
Nitrogen dioxide	Bacterial action in soil; oxidation of NO	2-2.5
Ammonia	Biological decay	4
Carbon monoxide	Oxidation of methane; forest fires; oceans	100
Ozone	Tropospheric reactions and transport from the stratosphere	20-60

Source: World Health Organization (1972)

and halocarbons (Francey, 1984). Table 1.1 lists for selected pollutants, estimates of background concentrations as derived by the World Health Organization (1972). Observations of pollutant concentrations within urban areas and about industrial complexes are usually many orders of magnitude above these levels (Lawlor, 1982).

Air pollution affects the well-being of humans, ecosystems and materials over areas ranging in scale from the local through regional and national and, more recently, to the global scale. With the advent of the Industrial Revolution the effects of air pollution were no longer restricted to the area local to the pollution source. The long range transport of acid gases in Europe and North America (United States National Research Council, 1983a) is not only due to the increased emission levels of recent times, but is also due to the construction of tall chimneys from which the primary pollutants are emitted. This strategy has converted a local pollution problem to one of international concern requiring international cooperation to ameliorate.

Examples of air pollution problems of global concern include the effects of the halocarbons upon the ozone layer (United States National Research Council, 1976) and the potential effects of increasing carbon dioxide levels (Bach, 1978; United States National Research Council, 1983b). It is hypothesised that the increased carbon dioxide levels brought about by the burning of wood and fossil fuels will lead to a global increase in temperature. Some of the effects of this increase in temperature are considered to be raised sea levels flooding many of the world's cities, and changes in the pattern of rainfall significantly altering agricultural production (United States National Research Council; 1983b).

In this thesis the problems examined have been restricted to the regional and local scales. Models are developed for area sources such as cities, and for large point sources which affect the air quality over large regions. At the local scale a model is developed to describe the dispersion of pollutants from roadway line sources. Obviously this model could be applied many times to represent a regional road network although this problem is not considered in this thesis.

In summary, air pollution arises from interactions within a complex system involving the emission of numerous pollutants from a wide range of sources. These pollutants are dispersed within the atmosphere where they may undergo transformation. Ultimately they affect ecosystems on a scale ranging from local to global. In the following section general approaches to modelling this complex system are considered.

1.3 Modelling atmospheric dispersion - a badly defined system

In a study of the long-range transport and deposition of pollutants Venkatram and Pleim (1985) considered that our understanding of this system is derived from two modelling approaches, namely the theoretical or 'reductionist' method and the empirical or 'holistic' method. This is true not only for long range transport of pollutants but for the study of atmospheric dispersion in general (Hanna, 1982a). Venkatram and Pleim (1985) consider that empirical models are important because of the difficulty in developing theoretically based models due to gaps in our understanding of the components of these models, lack of the extensive data sets required to run such models, and that these models generate responses which are not readily intelligible. They considered that empirical models whose structure is closely tied to observations were required to complement theoretical models.

While Venkatram and Pleim (1985) perceive modelling approaches at two extremes, they cite the work of Beck (1981) who describes the development of mathematical models as occurring over a wide spectrum. This spectrum varies from 'hard' systems such as electrical systems, to 'soft' systems such as social systems. The degree to which the behaviour of the system may be inferred a priori and planned experiments undertaken to verify the model formulation determines the 'hardness' of the system. Karplus (1976) and Vemuri (1978) describe this spectrum of models as ranging from white box systems (hard) to black box systems (soft). Air pollution systems are considered to fall within the grey area between these two extremes. Table 1.2 lists examples of systems lying within this spectrum.

Beck (1981) cites the work of Young (1978) who has suggested that natural environmental systems are difficult to analyse in mathematical terms because their underlying mechanisms tend to be 'badly defined'. The poor definition arises from the inability to conduct planned experiments on the system. Thus models lying within the grey area, as delineated in Table 1.2, may not necessarily move inexorably towards a white box model where a near complete understanding of the system has been reached. Hence uncertainty may remain a significant feature of the system (Young, 1978, ~~1981~~ 1982a).

Table 1.2: Spectrum of modelling activities.

Model character	Model description
Black box (or soft or empirical) model	Social systems Political systems
↑	Economic systems
	Physiological systems
	Air pollution systems
Grey area	Ecological systems
	Water pollution systems
	Industrial process control
↓	Aircraft control
White box (or hard or theoretical) models	Electrical circuits

Faced with badly defined environmental systems Young (1978, ~~1981~~ 1982a, 1982b) developed a methodology for the systematic analysis of systems based upon several considerations. While the methodology was specifically developed for dynamic systems some of the concepts are also relevant to systems that are modelled as static. These concepts are:

> (i) while the system is nominally complex its passively observed behaviour is often dominated by relatively simple linear or nonlinear relationships;
>
> (ii) that hypothetico-deductive procedures of the scientific method be used to establish those modes of behaviour consistent with the observations; and
>
> (iii) the identifiable modes of behaviour may not provide a total description of the system as further data collection may suggest that other modes are also significant.

This approach to modelling environmental systems has been applied successfully to water quality problems by Young (~~1981~~ 1982a), Hornberger and Spear (1980) and Jakeman et al. (1984) and has been applied to the study of air quality by Steele (1981) and Steele and Jakeman (1980). Most of these studies applied recursive methods of time series analysis developed by Young (1974, 1976), Young and Jakeman (1979, 1980), ~~Young et al. (1980),~~ Jakeman et al. (1980), and Jakeman and Young (~~1982a~~ 1979, ~~1982b~~ 1981, 1983).

In this study, while time series analysis methods have not been employed as the analytic tool, the general principles of modelling badly defined systems as outlined above have been adopted. Thus models developed here are simple models aimed at producing a parametrically efficient description of the air quality data. The models developed are not black box models but are simple descriptions of the physical dispersive system where many of the state variables and mechanisms are capable of clear interpretation.

1.4 Thesis outline

The thesis structure is as follows. Chapter 2 presents a survey of the relevant air pollution literature and introduces the principles for the development of the air quality models. In Chapters 3 and 4 the methods of parameter estimation and model identification necessary for the improved use of statistical models and for the development of statistical model components of hybrid models are examined. The subsequent three chapters form a series of investigations into the modelling of air quality. To demonstrate the generality of the approach models are constructed for the three key emission source regimes, namely for area, line and point sources. The problem of designing monitoring networks to provide maximum information return in terms of assessing compliance with air quality standards is considered in Chapter 8. A more detailed description of the individual chapters follows.

In Chapter 2 models capable of describing the distribution of pollutant concentration are examined. The two major approaches, deterministic and statistical modelling, and the advantages and limitations of each are considered. The combination of these two modelling approaches, the hybrid modelling methodology, is described here.

Numerous methods are availible for estimating the parameters of distributional models considered applicable to the study of air quality data. Chapter 3 examines various methods of parameter estimation using Monte Carlo experimentation. For each of the methods the bias, variance and more practical considerations such as the computational demands of the methods are examined. A method for generating approximate confidence intervals at percentiles of interest for each of the distributional models examined in the Monte Carlo studies is presented.

In Chapter 4, employing the methods of parameter estimation examined in Chapter 3, model identification procedures are considered. Based upon an examination of their performance a new procedure for the selection of appropriate distributional models for air quality data is presented. The model identification procedure is applied to air quality observations recorded in Melbourne, Australia.

In Chapter 5 a hybrid model combining a simple deterministic model and an identified distributional model is developed for predicting acid gas concentrations in the industrial city of Newcastle, Australia. Approximate confidence intervals were derived for model predictions and were found to yield a useful measure of model uncertainty.

The problem of predicting the dispersion of pollutants from roadway line sources is considered in Chapter 6. A brief review of models applied to roadway line sources and their performance is given. On the basis of model calibration and model validation exercises the hybrid model is found to yield considerably improved estimates, when compared with the deterministic model applied alone, of the upper percentiles of the distribution of pollutant concentrations.

Chapter 7 describes the development of a hybrid model for the dispersion of sulphur dioxide from point sources. The calibration of the deterministic component of the hybrid model is performed using the percentiles rather than time-wise matched pairs. The hybrid model produces estimates of pollutant concentration at 24-h, 8-h, 3-h, 1-h and 0.5-h averaging times. Compared with the deterministic model applied alone the hybrid modelling approach provides a significant improvement in the prediction of the upper percentiles of the distribution of pollutant concentrations observed about point sources. Approximate confidence intervals were derived for model estimates and were found to provide reasonable bounds for model uncertainty.

The need to design air quality monitoring networks to obtain the maximum return of information forms the basis of the work reported in Chapter 8. Here three approaches for increasing the spatial resolution of air monitoring networks based upon restricted data sets are considered. The importance of the application of a model identification procedure such as that developed in Chapter 4 is demonstrated. Where a second complete data set is available, an empirical quantile-quantile model may be applied. This modelling approach does not require the assumption of a distributional form for the air quality data. The empirical quantile-quantile model is found to provide the best estimates of the upper percentiles of the pollutant distribution.

Finally, in Chapter 9 a summary of the principal conclusions of the thesis and a discussion of future directions that this research may take are presented.

CHAPTER 2
MODELLING THE DISTRIBUTION OF AIR POLLUTANT CONCENTRATIONS

2.1 Introduction

This chapter examines the two major approaches to the problem of modelling the distribution of air pollutant concentrations. It should be noted that this thesis is concerned with pollutants which may be considered inert, or at least relatively inert, over the measurement time scale. The two modelling approaches considered can be broadly classified as deterministic and statistical modelling. Here the term 'deterministic' refers to models formulated using physical laws. These models yield a mechanistic description of the dispersion of pollutants within the atmosphere.

The origins of the deterministic approach to modelling dispersion are attributed to the work of Taylor (1915, 1921, 1927) who measured turbulent velocities in the horizontal plane using the widths of the traces produced by the wind speed and direction. This work was followed by full scale tracer experiments performed by Sutton (1932, 1934). Under near ideal conditions the first specifications of the cross wind and vertical spread of suspended material were obtained over a range of a few hundred metres from the source.

The term 'statistical' refers to those approaches where outputs are inferred from non-mechanistic non-physical relationships using statistical methods. Here models are developed to describe the distribution of pollutant observations. These models are descriptive and are not strictly applicable beyond the conditions existing when the data upon which they have been developed were collected. The stimulus for the development of these models can be attributed to the formulation of air quality standards written in terms of a pollutant concentration which must not be exceeded by more than a stated frequency. Larsen (1969, 1971, 1973, 1974), in an extensive analysis of data sets monitored in the United States, concluded that the lognormal model yielded a good fit to the distributions of all pollutants, over all averaging times and at all sites considered. This statement has attracted and continues to attract much attention and is considered to provide the empirical basis for the application of distributional models to air quality data. Attempts have been made to infer a priori the distribution that air quality data should follow. However, to date this approach has met with only limited success.

A second type of statistical modelling undertaken in this thesis relates the percentiles of the distribution of pollutant concentration using a simple linear relationship. This model, termed an empirical quantile-quantile model (Chambers et al., 1983), is of importance to the design of monitoring networks and the development of monitoring strategies where restricted data sets are collected.

In this chapter the deterministic and statistical modelling approaches will be examined with the view to identifying the strengths and the limitations of each in determining the distribution of pollutant observations. Further, the concept will be introduced that these two modelling approaches can be combined in such a manner so that the strengths of each approach are exploited while the weaknesses of each are attenuated.

2.2 Deterministic air quality models

In this section deterministic models for the dispersion of inert gases within the atmospheric boundary layer are critically examined. There are many reviews which cover the full range of air quality models including those by Lamb and Seinfeld (1973), Eschenroeder (1975), Johnson et al. (1976), Turner (1979), Simpson and Hanna (1981), Hanna (1982a) and Geraghty and Ricci (1984). The various theories of atmospheric dispersion are examined on the basis of how each predicts pollutant concentrations. Thus it is the different model treatments of turbulent diffusion which will be examined here. An understanding of the accuracy of air pollution model predictions is important in environmental management as this allows an assessment of the risk of exceeding air quality standards to be undertaken. Accordingly a discussion of the accuracy with which deterministic models predict pollutant concentrations and the likely limits to model accuracy is presented. The problem of assessing model performance is also considered as no single method for evaluating model performance is clearly best.

2.2.1 Gradient transport model

The set of equations that form the basis for the development of mathematical models for the dispersion of pollutants within the planetary boundary layer describe the motions of a viscous, compressible, Newtonian fluid in a rotating system. The equation of continuity, which is an expression of the conservation of mass is stated as

$$\frac{\partial}{\partial t}\rho + \frac{\partial u\rho}{\partial x} + \frac{\partial v\rho}{\partial y} + \frac{\partial w\rho}{\partial z} = 0 \qquad (2.1)$$

where ρ is the instantaneous density and u, v, w are the instantaneous velocity components describing the motion of the fluid in the x, y and z directions respectively, at the point (x,y,z) at time t and in a Cartesian co-ordinate system. Equation (2.1) may also be written in the form

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = \frac{-1}{\rho}\frac{d\rho}{dt} \qquad (2.2)$$

where the individual terms on the left side are usually at least two orders of magnitude larger than the right side. Consequently the assumption

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \qquad (2.3)$$

that the atmosphere is incompressible is a good and useful approximation (e.g. see Businger, 1982).

The differential equation which has become the starting point of most mathematical treatments of diffusion from sources is a generalization of the classical equation for the conduction of heat in a solid and is essentially a statement of the conservation of the mass of suspended material (Pasquill and Smith, 1983). Denoting the concentration by χ units of mass per unit volume of a fluid which is assumed incompressible we have

$$\frac{\partial \chi}{\partial t} = - \left[\frac{\partial (u\chi)}{\partial x} + \frac{\partial (v\chi)}{\partial y} + \frac{\partial (w\chi)}{\partial z}\right] \qquad (2.4)$$

The above equation constitutes only one of five coupled fundamental equations describing all aspects of the interaction of chemically active constituents in a fluid (Businger, 1982). However, since air pollutant concentrations only rarely exceed a few parts per million (ppm) by volume, the presence of air pollutants will produce insignificant changes to the heat balance of the atmosphere. An exception to this rule is the reduction in solar radiation intensity resulting from air pollutants over urban areas (Bach, 1971). In general, the assumption that meteorology remains unchanged is reasonable. This assumption allows equation (2.4) to be solved with the fluid velocities u, v, w considered independent of the concentration terms χ_i .

Unfortunately the complexity of turbulent flow is so formidable that even if we were able to describe its structure in detail, comprehension would be close to impossible (Businger, 1982; Pasquill and Smith, 1983). This has led to the development of the description of turbulent flow in terms of its statistical characteristics. It has been assumed therefore that the fluid motions can be separated into a slowly varying mean flow $(\bar{u},\bar{v},\bar{w})$ and a rapidly varying turbulent flow (u',v',w'). Hence the instantaneous velocity components in a rectangular co-ordinate system are given by

$$\begin{aligned} u &= \bar{u} + u' \\ v &= \bar{v} + v' \\ w &= \bar{w} + w' \end{aligned} \qquad (2.5)$$

Similarly the instantaneous concentration terms χ_i are themselves random variables

$$\chi_i = \bar{\chi}_i + \chi_i' \qquad (2.6)$$

Writing u, v, w and χ as the sum of a mean and eddy fluctuations as given in equations (2.5) and (2.6), expanding, averaging and rearranging equation (2.4) gives

$$\frac{\partial\bar{\chi}}{\partial t} + \bar{u}\,\frac{\partial\bar{\chi}}{\partial x} + \bar{v}\,\frac{\partial\bar{\chi}}{\partial y} + \bar{w}\,\frac{\partial\bar{\chi}}{\partial z} = \left[\frac{\overline{\partial(u'\chi')}}{\partial x} + \frac{\overline{\partial u'\chi'}}{\partial y} + \frac{\overline{\partial(w'\chi')}}{\partial z}\right] \qquad (2.7)$$

Physically the velocity components refer to an element of air passing through a specified point. In the Eulerian-space system the velocities are in principle specified at all positions in the field of flow at a given instant. In the system known as Lagrangian, the concern is with the variations in time of the velocity of a particular element, which is of course continuously changing its position. The analysis of atmospheric turbulence is usually concerned with the fluctuation recorded by a fixed instrument responding more or less rapidly to the relative motion of the air. This situation is customarily regarded as equivalent to the Eulerian-space description (Pasquill and Smith, 1983), the equivalence is based upon the hypothesis that the sequence of variations at a fixed point is statistically the same as the instantaneous spatial variation.

The gradient-transport approach assumes that turbulence causes a net movement of material down the gradient at a rate which is proportional to the magnitude of the gradient. The turbulent transfer of material in this manner is referred to as a simple diffusion process (Pasquill and Smith, 1983). By replacing the eddy flux terms by the simplest gradient-transport forms equation (2.7) becomes

$$\frac{d\bar{\chi}}{dt} = \frac{\partial}{\partial x}\left(K_x\,\frac{\partial\chi}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\,\frac{\partial\chi}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\,\frac{\partial\chi}{\partial z}\right) \qquad (2.8)$$

This equation allows for differences in the eddy diffusivities (K_x, K_y, K_z) in the component directions and also for the spatial variation of these diffusivities. If the K's are constant and independent of x, y or z then the resulting equation and type of diffusion implied are Fickian. In this case the distribution of material is of a Gaussian form with variances

$$\sigma_x^2 = 2K_x t = 2K_x x/u \qquad (2.9)$$

however, experimental studies have shown that the equivalent values of K vary systematically with the time of travel, with the position, and with the scale of the diffusion process (Pasquill and Smith, 1983).

The gradient-transport formulation may be expected to be the most successful when the diffusive action of the turbulence is effectively confined to scales small relative to the volume occupied by the suspended material. For vertical spread this condition is approached for ground level releases of pollutants. However, it is not met in the case of time-mean lateral dispersion from a continuous point source and for vertical dispersion from an elevated source.

2.2.2 Gaussian plume models

The Gaussian plume model is the most widely applied diffusion model (Hanna, 1982a; Turner, 1979). The Gaussian plume model takes its name from the assumed Gaussian form describing the vertical and cross-wind concentration profiles. As noted earlier the Gaussian plume formula can be derived from equation (2.8) providing that the turbulence is homogeneous and stationary and only a point source is considered. The Gaussian expression is (e.g. Hanna, 1982a)

$$\chi\ (x,y,z) = \frac{Q}{2\pi u\sigma_y\sigma_z} \exp\ (\frac{-y^2}{2\sigma_y^2})\ [\exp\ (\frac{-(z-H)^2}{2\sigma_z^2}) + \exp\ (\frac{-(z+H)^2}{2\sigma_z^2})] \qquad (2.10)$$

where Q is the source strength (mass emission rate), u the mean wind speed, σ_y and σ_z the standard deviations in concentration in the crosswind (y) and vertical (z) directions respectively, and H is the effective height of emission. The wind flow is assumed to be in the x-direction.

The application of equation (2.10) is based upon the following assumptions that:

(a) both the emission rates and meteorological conditions have attained a steady state;

(b) a constant windspeed and wind direction, valid for the entire region, may be specified;

(c) no absorption or generation by the ground occurs;

(d) there is no inversion layer;

(e) the diffusivities in the vertical and cross-wind directions vary only with downwind distance and are constant in the diffusion domain; and,

(f) the pollutant does not undergo chemical reaction.

Observations of passive plumes have confirmed that the Gaussian form is a satisfactory description for the cross-wind distribution, at least for ensemble averages (Pasquill and Smith, 1983). In spite of the above simplifying assumptions necessary for the application of the Gaussian plume model, this model has seen widespread application (Simpson and Hanna, 1981; Turner, 1979). This model, in various forms for point area and line sources constitutes the basis of the UNAMAP (User's Network for Applied Modeling of Air Pollution) developed by the United States Environmental Protection Agency as an aid to the development of air quality management strategies.

The Gaussian plume model has also been modified by Davis and Metz (1978) for the special case of particulate matter to allow for surface deposition and reflection. More recently Hanna et al. (1984) modified the Gaussian plume model for application in complex terrain. The Gaussian plume formula as stated in equation (2.10), with suitable integration may be applied to line and area sources. The specific details of the Gaussian plume models developed in this study for line and point sources are presented in Chapters 6 and 7. The models developed in these chapters employ the most recent improvements to the Gaussian plume model (see Hanna, 1982b) in such areas as wind speed evaluation, plume rise calculations and the estimation of σ_y and σ_z. The application of these improvements is limited only by the availability of the data necessary for their implementation.

2.2.3 Rollback models

The rollback model was developed to provide a simple method by which source pollutant emissions could be assessed (Chang and Weinstock, 1975). The original model assumes that pollutant concentrations are directly proportional to emissions according to some simple relationship. Thus the emission control requirements are presumed proportional to the amount by which the peak pollutant concentration exceeds the standards. The nonlinearity of the atmospheric processes limits the usefulness of the rollback model to a role in which a first rough estimate is made of the emission controls required. The simplest form is of the type

$$\chi = ke + b \qquad (2.11)$$

where χ is the pollutant concentration due to emissions at a rate e, with b being a measure of the background pollutant concentration and k the constant of proportionality. All the effects associated with the meteorology, the distribution of sources and all other factors are included in k. De Nevers and Morris (1975) define k in terms of the highest concentration

$$k = (\chi_{max} - b)/e \qquad (2.12)$$

where χ_{max} is the highest pollutant concentration in the region of interest. The allowable emission rate e_{max} is then

$$e_{max} = \frac{e\ (\chi_{std} - b)}{(\chi_{max} - b)} \qquad (2.13)$$

where χ_{std} is the air quality standard required for the pollutant being considered.

A more general form of the simple rollback model allowing for a range of source and receptor interactions is given by

$$\chi_i = \Sigma\ k_{ij}\ e_j + b \tag{2.14}$$

where χ_i is the concentration at receptor i, e_j is the emission rate at source j, and k_{ij} is the source-receptor interaction for source j and receptor i. Peterson and Moyers (1980) have extended the above model to the case where continuous measurement of ambient concentrations and emissions are available and recorded over time intervals corresponding to air quality standards.

Georgopoulos and Seinfeld (1982) examined these rollback calculations and recommended the replacement of χ_{max} and χ_{std} with the expected mean values, denoted $E(\chi_{max})$ and $E(\chi_{std})$ respectively where $E(\chi_{std})$ represents the expected mean concentration from a distribution in which χ_{std} is an extreme. Applying the expected values in rollback calculations allows for the conservation of mass of non-reactive pollutant. Georgopoulos and Seinfeld (1982) then considered whether the concentration χ_{max} would increase linearly with emissions as does the expected value $E(\chi_{max})$. They found that linearity holds for the particular case where pollutant concentration is lognormally distributed and meteorological conditions remain unchanged. In particular the geometric standard deviation must remain unchanged. There is of course ample evidence indicating that these conditions are rarely met (Simpson et al, 1985). Thus rollback models are only useful at the initial stages or as screening models with which a crude indication of future trends may be determined.

2.2.4 Box models

A useful evaluation of the effects of a large area source may be made using the simple box model described by Gifford and Hanna (1971, 1973), referred to as the ATDL (Atmospheric Turbulence and Diffusion Laboratory) model. This model has seen widespread application (Eschenroeder, 1975; Pasquill and Smith, 1983). The ATDL model is applicable to urban area sources in which the emissions are assumed uniform over grid squares that may vary in size from 1- to 10-km square. For a grid pattern with uniform source strengths in each square, the ground level concentration χ_o in a grid square is given by

$$\chi_o = (2/\pi)(\Delta x/2)^{1-b}\,[ua\ (1-b)]^{-1} \cdot \{Q_o + \sum_{i=1}^{N} Q_i\ (2i + 1)^{1-b} - (2i - 1)^{1-b}\} \qquad (2.15)$$

where N is the number of upstream grid squares contributing to the concentration in the grid square under consideration (designated by the subscript o), a and b are parameters dependent on atmospheric stability, Δx is the size of the grid square, u is the average horizontal windspeed assumed to be in the x-direction, and Q (i = 0, 1, 2, ..., N) are the area source strengths for each grid square.

When the spatial variation in the source strengths are smooth slowly varying functions the above equation may be simplified (Hanna, 1971) to

$$\chi_o = \frac{C\ Q_o}{u} \qquad (2.16)$$

where C is a constant that depends on atmospheric stability. Equation (2.16) follows from (2.15), only for smooth area source distributions in which the terms involving Q_i $(i \neq o)$ in (2.15) are significantly less than the Q_o term. Benarie (1976) has shown that this equation is widely applicable.

The ATDL model has been shown to yield predictions that compare favourably with those given by four more complex models (Hanna, 1971) and has been demonstrated to be applicable to a wide range of urban environments including Frankfurt (Hanna and Gifford, 1977), Canberra (Daly and Steele, 1976) and Milan (Gualdi and Tebaldi, 1982).

Benarie (1978) examined the box model and revised the usual formulation to include a parameter for the exponent of the windspeed. His form of the box model is then

$$\chi = CQu^{p} \tag{2.17}$$

where p is the windspeed exponent which may vary between 0 and -1. Benarie considered that only where advective mixing dominates would the -1 power, as stated in equation (2.16), be correct. In the case of predominant convective mixing, Benarie states that advection would not change the concentration and thus windspeed should have an exponent of zero. Benarie found from a review of experimental studies that the value of the exponent ranged between -0.2 and -0.5. Benarie also speculated that the exponent p may be a climatological characteristic for any given city and may vary with season.

Daly and Steele (1976) and Simpson et al. (1983) have relaxed the assumption implied in equation (2.16) that the windspeed and pollutant concentration are inversely related as matched pairs of observations. Instead it is assumed that opposite percentile values of windspeed and air pollution distributions are related by a simple inverse relationship, stated as (Simpson et al., 1983)

$$\chi_p = \frac{K}{u_{100-p}} \tag{2.18}$$

where χ_p is the air pollution concentration corresponding to the p-percentile ordinate of the air pollution cumulative frequency distribution, u_{100-p} is the windspeed corresponding to the (100-p)-percentile ordinate of the windspeed cumulative frequency distribution and K is a constant. The constant is derived from the relationship between χ_p and u_{100-p} for each sampling station under consideration over some percentile range for which K is approximately constant. Simpson et al. (1983) use the medians to estimate K so that

$$K = u_{50}\chi_{50} \tag{2.19}$$

Knox and Lange (1974) and Benarie (1976) suggested a similar measure and noted that the constant K derived in this manner requires no direct knowledge of the source strength. Simpson et al. (1985) observed for total suspended particulates and acid gas observations that $\chi_p u_{100-p}$ is reasonably constant over the 30-70 percentile range. Thus for their data at least ~~that~~ the model given by equation (2.18) is a good representation of the relationship of the statistical distributions of windspeed and pollutant concentration for at least the 30-70 percentile range.

The model as given in equation (2.18) was combined with the assumption of a lognormal distribution of pollutant concentrations and windspeed data to yield estimates of the entire distribution of pollutant concentration (Simpson et al., 1983). The p-percentile concentration is easily found from

$$\chi_p = \frac{K}{\alpha_u} (\beta_u)^{z_p} \qquad (2.20)$$

where β_u is the geometric standard deviation, α_u the geometric mean (median) of the windspeed data and z_p is the standard variate corresponding to the p-percentile. This model, equation (2.20) with (2.19), has been applied by Simpson and Jakeman (1985) to forecast worst case pollution scenarios for acid gas and suspended particulates due to urban industrial development. Thirty years of windspeed data were parameterised to obtain estimates of α_u and β_u for each year in order to obtain a range of extreme values in (2.20) for a given K.

2.2.5 Deterministic model performance

Several discussions have recently appeared on the accuracy attainable in the prediction of air pollution concentrations (Hanna, 1982a; Benarie, 1976, 1982; Pasquill and Smith, 1983; Venkatram, 1984). In principle, were the boundary and initial conditions completely specified, the solution of the equations of motion would describe all the details of turbulent flow and hence air pollutant concentration. However this is impossible in practice due to the nature of turbulent flow. We treat this motion as a mean flow with which can be associated an infinite set of unresolved residual turbulent flows. Therefore the best an air quality model can be expected to produce are estimates of average concentrations (Venkatram, 1983) and a model estimate can be expected to differ from the corresponding observation.

This conclusion is supported by the study of Hanna (1982c) who from an analysis of ~~a~~ meteorological and pollutant data sets examined the effect of "natural variability". Hanna considered the variation inherent in pollution concentrations based upon the division of pollutant observations into 18 wind direction classes, 10 wind speed classes and 7 stability classes. The data examined were hourly carbon monoxide and sulphur dioxide data recorded at 25 stations in St Louis for all hours of 1976. Hanna (1982c) concluded that the natural variability of hourly average concentrations in St Louis for given meteorological and source conditions is typically a factor of two.

It could be argued that the limitations encountered in the application of deterministic models are due to the lack of meteorological information, particularly in respect of special factors which are required for the full application of the more complex models. However, it has been recognized that the cost of the required observational programmes would more than likely outweigh any benefits achieved from the increased reliability of pollutant concentration estimates (Pasquill and Smith, 1983). The proposal for increased data collection has been further criticized on the basis that the underlying dispersion relations apply at best only to idealized situations of air flow and topography which are rarely of practical significance particularly when determining the more adverse conditions of dispersion (Benarie, 1982; Pasquill and Smith, 1983).

It is not surprising then, that comparisons of simple and complex modelling methodologies have shown that model performances are similar (Simpson and Hanna, 1981; Benarie, 1976, 1982). Pierce (1984) in a study of point source Gaussian plume models used to estimate hourly sulphur dioxide concentrations, noted that the more sophisticated models did not appreciably outperform the routinely applied models. Given the greater input data requirements and the usually significantly larger computational requirements of the more complex models, a simple model is usually preferred. Hanna et al. (1984) noted that the computational demands of a model for hourly pollutant observations recorded over two years limited the complexity of the model that could be developed. The models developed in this thesis are based upon the Gaussian plume assumption which allows an analytical approach. On the other hand, the K diffusion models require numerical solution employing methods such as finite elements or finite differences which themselves may introduce error (Chock, 1985a; Pasquill and Smith, 1983).

For most practical applications the accuracy of air quality models for ensemble averages is not expected to be less than several tens percent, while for comparisons with individual observations factors of two or more may be expected (Benarie, 1976, 1982; Hanna, 1982b; Pasquill and Smith, 1983; Simpson and Hanna, 1981). For the point source Gaussian plume model Nieuwstadt (1980) found that an accuracy of a factor of 2 for mean concentrations, and for the upper percentiles a factor 4, could be expected. Turner and Irwin (1982) when using a point source Gaussian plume model to predict the second highest 3-h and 24-h average sulphur dioxide concentrations observed considerable scatter between model predictions and observations. For the 37 data sets they examined for 24-h periods 68% of model predictions were within a factor of 2 while for 3-h periods 84% were within a factor of 2. No significant biases in model estimates were revealed. Venkatram (1984) in a theoretical analysis of the uncertainty of model predictions of 1-h average concentrations considered that 25% of the observations would lie outside a factor of two of the maximum predicted concentration.

In summary the level of uncertainty expected from simple deterministic models would be in the order of a factor of 2 for ensemble means and larger for the more extreme pollutant concentrations. Finally Pasquill and Smith (1983) note that, "It remains to be seen whether any significant improvement in this capability will ensue as a consequence either of the ever-increasing sophistication in the studies of atmospheric flow or of the continuing elaboration in mathematical modelling techniques". This view is supported by the theoretical considerations discussed by Benarie (1976, 1982) and Venkatram (1984), and from the analysis of observational data by Hanna (1982c).

2.2.6 Assessing model performance

The assessment of the performance of air quality models cannot be undertaken simply by application of a few clear-cut criteria. This is because most air quality models are essentially empirical in nature. Each model needs to be assessed on its own merit while keeping in mind the purpose for which the model was developed. Model performance and model validity should be distinguished. In general, the evaluation of model performance centres upon the comparison of the observations with model predictions whereas model validity refers to the fundamental correctness of the model formulation. Clearly when considering the behaviour of air quality models many aspects must be examined, such as the magnitude of model predictions and the spatial and temporal resolution.

Bencala and Seinfeld (1979) reviewed methods for assessing air quality models and categorized the methods under the headings of the analysis of residuals, the analysis of trends and the analysis of indices of air quality. A fourth category, other analyses, included the chi-square test and spectral analysis. The analysis of residuals consisted essentially of the evaluation of mean and root mean square errors and included plots of the residual errors versus time. Residual errors were also plotted against observed and predicted concentrations and also for each monitor location. The analysis of trends comprised the evaluation of the correlation coefficient, regression analysis and included scatter plots. The indices of air quality centred on comparing high predicted and observed pollutant concentrations. The frequency distribution of the predicted and observed concentrations were also plotted.

Hayes (1979) also reviewed the criteria which could be used to evaluate air quality model performance. Hayes showed that the criteria adopted depended largely on what the model was designed to predict. More recently the United States Environmental Protection Agency recommended three statistical performance measures (Fox, 1981). These were bias, average gross error and the correlation coefficient. The evaluation of the bias and average gross error were normalized to eliminate differences in magnitude by division with the observed concentration. In addition, it was suggested that plots of the observed and predicted cumulative distribution functions be prepared.

Several authors have noted the importance of graphical presentations to assist in the analysis of model performance. These methods should be considered complementary to numerical assessments (Eschenroeder, 1975; Simpson and Hanna, 1981). In particular the study of Anscombe (1973) is cited where four data sets exhibiting very different graphical forms produced the same correlation coefficients. Chambers et al. (1983) provided an extensive review of the many graphical methods available for data analysis and demonstrated the importance of graphical methods for various applications and the advantages of particular methods.

In the following chapters the performance of the models developed are assessed using, where appropriate, a range of numerical indicators including those described by Fox (1981) and Bencala and Seinfeld (1979). Graphical methods are also used extensively in order to both support the numerical analyses and reveal more detailed model behaviour, especially with regard to the spatial and temporal variations.

2.3 Statistical models for air quality concentrations

In the preceding section the range, value and limitations of deterministic models in predicting the distribution of air quality observations were discussed. The important advantages of this approach were identified as (a) the link between the driving variables such as emissions and meteorological conditions and (b) the ability to predict ensemble averages well. A major limitation of this modelling approach was the inability to predict the upper percentiles of distribution of pollutant concentrations. In this section models which describe the entire distribution of air quality observations are examined.

These models are phenomenological in that they only seek to describe the observations rather than relate the atmospheric and emission variables to the observations. A distributional model is a description of the behaviour of a random variable. The distribution function determines how the cumulative probability is distributed over the possible values that a random variable may take. Air quality observations are considered to be random variables as they represent the result of fluctuations of both the meteorological conditions governing dispersion, in particular the turbulent flow, and the factors affecting pollutant emission rates, such as a change in human activities. Based upon assumptions requiring that the pollutant observations be independent and identically distributed, a probability model may be constructed.

The major advantages of the statistical model are that (a) it provides a simple representation of the entire distribution of pollutant concentration; (b) it allows interpolation between, and extrapolation from, the observed concentrations; (c) it may be readily plotted to yield a complete pictorial representation; (d) it provides the basis for further statistical analyses such as the evaluation of confidence intervals and the analysis of trends; (e) inferences regarding the nature of the underlying physical processes may be generated; and, (f) estimated pollutant concentrations may be directly compared with air quality standards.

By contrast the major disadvantages of statistical models are that they are non-causal and thus cannot readily be related to the driving variables of the air pollution system such as emissions and meteorology. As a consequence statistical models are restricted to the conditions under which they were developed.

As air quality standards form the basis upon which many air quality management decisions are formulated the nature of these standards are considered next. It will be shown that while air quality standards may change in regard to the actual level specified, the statement of these standards in probabilistic terms is likely to remain unchanged.

2.3.1 Air quality standards

The formulation of air quality standards represents the result of the complex interaction of scientific, economic, social and political considerations (Newill, 1977). Scientific considerations include an examination of the health effects (Ferris, 1978; Wyzga, 1978; Chamberlain, 1983) biological effects (Larsen and Heck, 1976, 1984) and physical effects of air pollutants whereas the economic, social and political considerations (Fisher, 1981; Brady et al., 1983) determine what control strategies are affordable and what level of control is socially and politically acceptable. Thus, air quality standards are judgemental and represent an expression of public policy. Air quality standards are subject to review based upon a changing understanding of the effects of air pollutants and as a result of changing public perceptions and expectations. Air quality standards are often formulated in terms of long and short term goals. For example ambient air quality standards recently introduced in China (Siddiqi and Chong-Xian, 1984) were established at three levels. The first level represented the ideal standard at which no harmful effects occur, the second level was believed to be the threshold at which effects become detectable and the third level was considered necessary to protect people from acute or chronic poisoning and to protect animals and plants (except sensitive ones). Standards at

Table 2.1: World Health Organization recommended long-term goals.

Pollutant	Limiting level (μgm^{-3})	
Sulphur oxides (a)	Annual mean	60
	98% of concentrations below[b]	200
Suspended particulates(a)	Annual mean	40
	98% of concentrations below[b]	120
Carbon monoxide	8-h average	10000
	1-h maximum	40000
Photochemical oxidants	8-h average	60
	1-h maximum	120
Nitrogen dioxide	1-h average not to be exceeded more than once a month	200-340

(a) Values for sulphur oxides and suspended particulates apply only in conjunction with one another.

(b) The permissible 2% of observations over this limit may not fall on consecutive days.

the first level are determined by the central government of China, whereas standards at the second and third levels are designated by local governments based upon the activity taking place in those areas.

The World Health Organization has formulated air quality criteria based upon analyses of the relationship between ambient pollutant concentrations and the associated adverse effects. Air quality goals are the concentrations of particular pollutants which are believed to represent a safe level so that seriously adverse effects on human health and welfare, ecological systems and materials do not occur. Table 2.1 presents the long-term goals recommended by the World Health Organization for sulphur oxides, suspended particulates, carbon monoxide, photochemical oxidants and nitrogen dioxide. The majority of air quality goals listed in Table 2.1 represent a probability statement of the form that the pollutant concentration recorded over averaging time t is not to exceed $x\ \mu gm^{-3}$ with more than probability p. The development of air quality standards in this manner reflects the uncertainty associated with the prediction of dispersion of pollutants within the atmosphere and the desire to reach a goal that is practically attainable.

The United States standard for ozone was revised to include the concept of expected exceedances (Curran and Cox, 1979). It is this statement of air quality standards in terms of an allowable number of exceedances which has stimulated the application of distributional models to the analysis of air quality data. These models readily yield the expected number of exceedances which can be directly compared with the air quality standard. Even where an air quality standard is stated in terms of a concentration which is not to be exceeded, the expected probability of exceeding this level should still prove to be a particularly useful measure of the performance of control strategies.

2.3.2 Applying statistical models to air quality data

Many studies of the application of distributional models to air quality data analysis have been undertaken. Reviews of the application of statistical distributions include those by Pollack (1975), Bencala and Seinfeld (1976) and Georgopoulos and Seinfeld (1982). A wide variety of distributional models have been employed to describe air quality data. Larsen (1969, 1971) performed a comprehensive analysis of seven pollutants in eight cities and suggested that the two-parameter lognormal distribution was suitable for all pollutants over all averaging times. Surman et al. (1982) applied the Larsen model to carbon monoxide and total suspended particulate data in Brisbane, Australia. Lynn (1974) compared the two- and three-parameter gamma and lognormal distributions, the normal distribution, the four-parameter beta distribution and the Pearson distributions and noted that the two-parameter gamma and lognormal models were preferred over the three-parameter models. Lynn considered that this was due to the sensitivity of the method of moments fitting procedure to the extreme upper tail. Pollack (1975) also considered the two-parameter Weibull distribution as applicable to air quality data along with the lognormal, gamma and Pearson distributions.

Extreme value distributions have been considered by Singpurwalla (1972), Roberts (1979a, 1979b), Horowitz (1980) and Chock (1984). Giugliano (1985) describes an empirical model for predicting extreme values by relating the upper percentiles to average concentrations. The estimation of the parameters of this distribution requires that a sample of the extreme values be available. In practice, and certainly within the Australian context, these data may not be available or may only consist of few observations. The models developed also suffer from the same limitation applicable to all statistical models of the distribution of air quality data in that they remain essentially a description of the data upon which they were developed.

Ott and Mage (1976) and Mage and Ott (1978) suggested a censored three-parameter lognormal distribution as applicable to air quality data. The additional parameter to that of the usual two parameters was suggested in order to straighten plots of air quality observations on lognormal probability paper. This model estimates the third parameter of the distribution using graphical techniques where the value of this parameter is adjusted until a straight line results when plotting pollutant observations on lognormal graph paper. Difficulties with the analysis of a large number of data sets and with the estimation of the confidence limits of the model parameters limits its usefulness.

Bencala and Seinfeld (1976) examined the two and three-parameter lognormal, Weibull and gamma distributions and found that the three-parameter lognormal model was, overall, the best model, although the other models yielded better results in some cases. The superior performance of the three-parameter lognormal model over that of the two-parameter model was attributed to the added flexibility afforded by the extra parameter. The results of Lynn (1974) show that the addition of the third parameter will not always improve the fit to air quality data.

Cats and Holtslag (1980) applied the lognormal model to hourly sulphur dioxide data recorded in the Netherlands. They found that the two-parameter lognormal model provided a satisfactory estimate of the 98-percentile concentrations. Other models were not examined. Berger et al. (1982) studied 24-h average sulphur dioxide concentrations recorded in the Gent region of Belgium. It was observed that a two-parameter gamma distribution was a better representation of the whole ensemble than the usual lognormal. Tsukatani and Shigemitsu (1980) compared the performance of the two-parameter lognormal model and the Pearson system of distributions to fit hourly sulphur dioxide data recorded along Osaka Bay, Japan. While the Pearson system proved to be more flexible, the lognormal distribution was found to be a useful representation of air quality data except about isolated sources.

The exponential distribution has been proposed as applicable to air quality data recorded near isolated point sources (Gifford, 1974). Gifford (1974) cited the tracer studies of Barry (1971) and Scriven (1971) to support the results of the theoretical analysis indicating that an exponential distribution of pollutant observations would result from dispersion from an isolated source. Curran and Frank (1975) also believed that an exponential or Weibull distribution may in general yield a better fit to air pollutant data. More recently Simpson et al. (1984) and Jakeman and Simpson (1985) have successfully applied the exponential distribution to sulphur dioxide concentrations recorded about an isolated point source.

Distributional models have also been applied in areas related to the study of air quality. Mage (1980) employed the Johnson S_L and Johnson S_B models to the distribution of windspeeds. A comparison of the performance of other models was not undertaken. An exponential distribution produced excellent correspondence with observed precipitation chemistry ion concentrations (Pack, 1982). Again, however, a comparison with other candidate distributional models was not reported.

2.3.3 Developing distributional models for air quality observations

While much attention has been given to the application of distributional models to air quality data, as evidenced in the preceding section, the same cannot be said about the process by which distributional models are developed. It is important to examine this process as the range of models which can be applied to the analysis of air quality data is large. No one model can be clearly identified as best for all applications, and at present no one distributional model exists which could be selected a priori to represent the distribution of air quality observations (Georgopoulos and Seinfeld, 1982).

The process of constructing a distributional model for air quality data may be divided into four stages. These are: (a) the selection or identification of the appropriate distributional model; (b) the estimation of the model parameters by fitting the model to the observations; (c) the estimation of the uncertainty associated with parameter estimates; and, (d) estimates of confidence intervals for the percentiles of the distribution, termed interval estimation.

Initially the identification of a suitable distributional model for air quality data was undertaken by applying graphical techniques. Larsen (1969, 1971) identified the two-parameter lognormal model as applicable to air quality data and Ott and Mage (1976) identified and estimated one parameter of a three-parameter model using graphical techniques. By comparison mathematical goodness-of-fit tests have seen relatively limited use. The chi-square test was applied by Tong and De Pietro (1977) to sulphate data which confirmed the hypothesis of lognormality for nearly all data sets examined. Kalpasanov and Kurchatova (1976) applied the Kolmogorov test to pollutant data recorded in Sofia, Bulgaria and found that the hypothesis of lognormality was rejected in most cases. Bencala and Seinfeld (1976) compared the performance of several distributional models using a least squares fit criterion, and concluded on the basis of the data examined, that overall a three-parameter lognormal model gave the best fit. Ott et al. (1979) used the chi-square statistic, Kolmogorov statistic and value of the log likelihood function to test the validity of the lognormal model and to assess methods of parameter estimation. They found that if the problems of statistical independence are ignored, then the two-parameter lognormal model would be rejected with a high degree of confidence for all the carbon monoxide data sets considered. Holland and Fitz-Simons (1982) developed a computer program to fit six distributions with nearly all having three or four parameters using the method of maximum likelihood. Their program evaluated six goodness-of-fit criteria including the chi-square, absolute deviation and Kolmogorov statistics, but no comparison of the performance of the goodness-of-fit statistics was presented.

In Chapter 4 the problem of identifying a suitable distributional form for air quality data is examined in detail. Using Monte Carlo experimentation the ability of several goodness-of-fit tests to select the 'best' model is examined. The problem of selecting from a range of alternative models is considered and a procedure for increasing the probability of selecting the best model developed. Using this model identification procedure a large urban air quality data set is examined.

The estimation of the parameters of a distributional model may be achieved using numerous techniques (Johnson and Kotz, 1970). The two most commonly applied methods are the method of moments and the method of maximum likelihood. Very little attention has been given to the effect that the method of evaluation of the parameters of the distributional form has upon the estimates of the percentiles of the distribution. Mage and Ott (1984) on the basis of 100 Monte Carlo experiments concluded that the method of maximum likelihood provided the best parameter estimates for the two-parameter lognormal distribution when compared with the method of fractiles and method of moments. No studies have been reported in the air pollution literature of other methods of parameter estimation for the lognormal distribution, or of methods of parameter estimation for other distributions. In Chapter 3 several methods of parameter estimation for the two-parameter lognormal, two-parameter gamma, two-parameter Weibull and the one-parameter exponential distributions are examined. The assessment of model performance is based upon the ability of each model to predict the upper percentiles of the distribution.

The estimation of confidence intervals, exact or approximate, for model parameters and the resulting percentile estimates, has to date received little attention. A simple procedure for the estimation of confidence intervals is not available and studies of the distribution of air quality data do not report confidence intervals for model predictions. In the chapter following, a simple method for obtaining approximate confidence intervals for various sample sizes and parameter values is developed. In Chapters 5, 6 and 7 the models developed include a submodel which approximates the confidence intervals associated with estimates of particular percentiles.

2.4 The hybrid modelling approach

The methodology employed in this thesis combines the two areas of air quality modelling examined previously to yield the hybrid deterministic/statistical model. The hybrid modelling approach was developed to allow estimation of the entire distribution of air quality data, in particular the upper percentiles, to be reliably estimated from major causal variables and to provide approximate confidence intervals for these estimates. This form of model output is desirable as it may be compared with air quality standards. The measure of uncertainty associated with the model estimates allows an assessment of the risk of exceeding an environmental standard to be determined. Thus the hybrid modelling approach is particularly suited to the problems of the management of airsheds (including monitoring strategies) and the assessment of pollution control strategies.

The hybrid model offers numerous advantages over the application of deterministic or statistical models alone, these are:

(i) it accepts the inevitability of uncertainty in modelling air quality and seeks to quantify this uncertainty in model predictions;

(ii) the deterministic model component relates estimates of air pollution levels to causal variables such as meteorological and emissions data;

(iii) the statistical component using only the deterministic model output within its range of greatest reliability (usually about the mean concentration) provides estimates of the entire distribution of pollutant concentrations;

(iv) the statistical component allows approximate confidence intervals for any percentile of the distribution of pollutant concentration to be determined;

(v) the approach is applicable to a wide range of pollutant emission source types and hence to the many problems of air quality management;

(vi) the hybrid modelling approach is sufficiently general that it is not restricted to a particular deterministic or statistical model component and will allow new model components to be readily incorporated within existing hybrid models as our understanding of the deterministic and statistical descriptions of air quality data improves; and,

(vii) using the deterministic component, with its link to the variables governing dispersion, the natural variability due to long-term meteorological change and the effect of emission control strategies such as changing stack heights and pollutant emission levels, can be incorporated into the analysis of impacts and through the statistical component these effects can be compared with air quality criteria.

While the hybrid modelling approach offers the above advantages, the following limitations should also be noted:

(i) the deterministic model component must be capable of predicting a percentile range of the distribution of pollutant concentration reliably;

(ii) the distributional form assumed as the statistical component of the hybrid model must remain consistent as input variables change; and

(iii) the hybrid model will not predict when a particular pollutant concentration will occur but rather with what frequency a pollutant concentration will be exceeded.

Although the air quality literature currently abounds with descriptions of deterministic and statistical models which have been applied to predict the distribution of air quality observations few studies report models combining the two approaches. Some early studies have been found indicating that deterministic and statistical models might be usefully combined. At the end of a lengthy review of air quality models Eschenroeder (1975) considered that deterministic model output did not match user needs and speculated that analytical linkages could be made between observed frequency distributions and computed (deterministic) model results. Eschenroeder suggested estimating the parameters of a lognormal model using the distribution of the deterministic model output.

Benarie (1976) linked the distribution of windspeeds and pollutant observations through a simple inverse relationship between percentiles. The constant in this model was obtained as the multiple of the medians of the windspeed and pollutant data sets as had been suggested by Knox and Lange (1974) from their work with hourly averaged carbon monoxide data recorded in San Francisco. This proportionality factor was found to be nearly constant for all percentiles.

Liu and Moore (1984) used random inputs instead of inputs matched by time with a point source Gaussian plume model. They found that uncoupling the time linkage in the model input had no systematic effect on the predicted cumulative frequency distribution of concentrations.

Only recently has the hybrid approach been applied to the study of air quality data. Most notably Simpson et al. (1983) combined the ATDL model with Larsen's statistical model to yield estimates of pollutant concentration. This model was limited to the assumption of a lognormal distribution of both the air quality and windspeed data sets. However, where these assumptions were met, the upper percentiles of the distribution of air pollutant data were predicted within the accuracy usually observed only for mean concentrations. Estimates of the uncertainty associated with model predictions were not developed in this study, partly because the Larsen model is not readily amenable to this analysis. Simpson and Jakeman (1984) extended their analysis to estimate the effect of fluctuations in long-term meteorology on observed maximum acid gas levels. Simpson and Jakeman (1985) demonstrated the application of this model as an air quality management tool which could be used in the planning of industrial development.

In this thesis the hybrid modelling methodology is developed and demonstrated for the three key emission source types: area, line and point sources. For each application a deterministic model component was selected and the statistical model component identified from amongst a range of alternatives. The deterministic model component provides estimates of the pollutant concentration but only about the median concentration. Using a percentile range about the median the parameters of the statistical model component and associated uncertainty were estimated. With these data the entire distribution of pollutant concentration and approximate confidence intervals at any percentile, may be inferred. Figure 2.1 illustrates the major steps in the application of the hybrid modelling methodology to area, line and point pollutant emission source types.

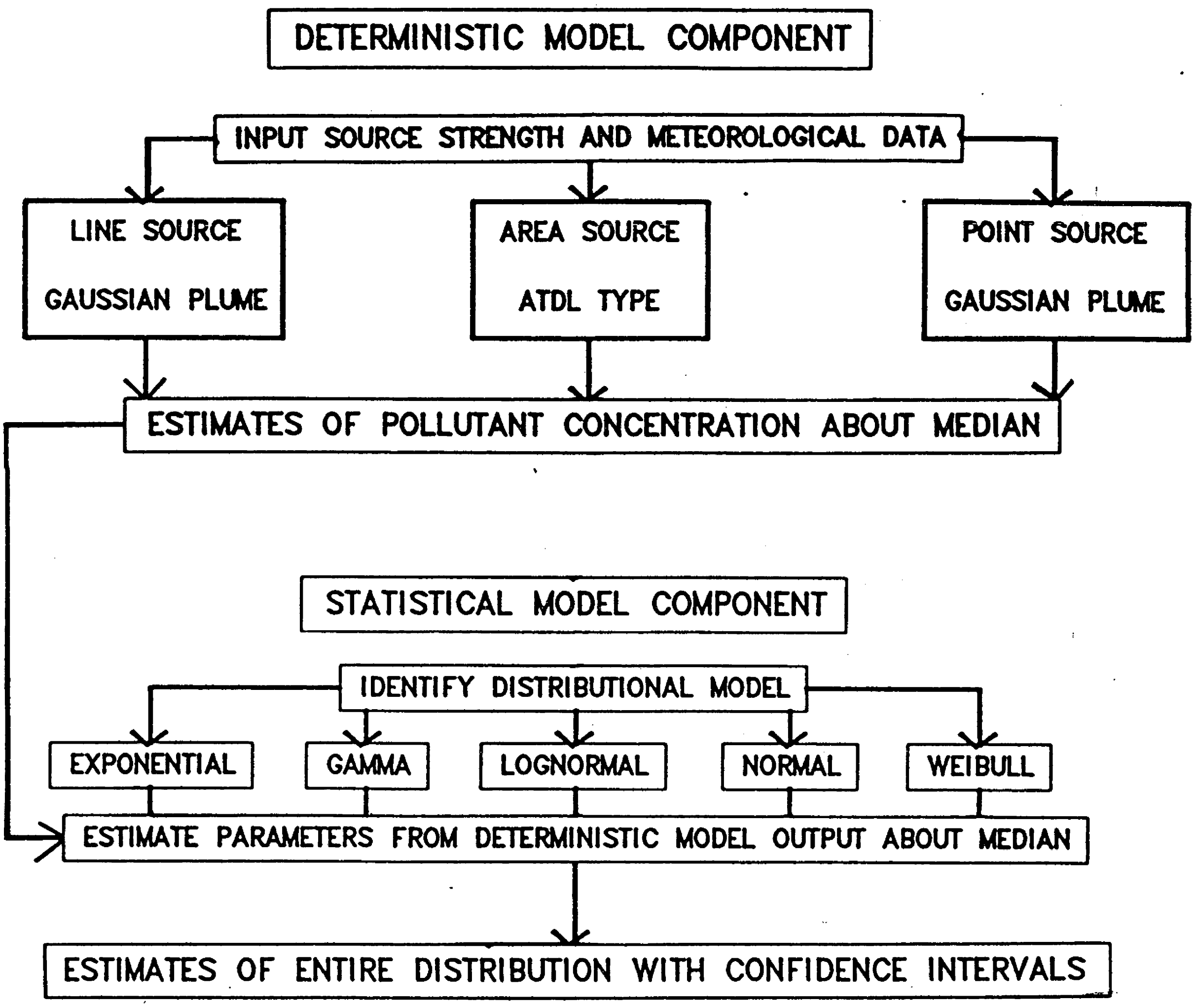
DETERMINISTIC MODEL COMPONENT
INPUT SOURCE STRENGTH AND METEOROLOGICAL DATA
LINE SOURCE GAUSSIAN PLUME
AREA SOURCE ATDL TYPE
POINT SOURCE GAUSSIAN PLUME
ESTIMATES OF POLLUTANT CONCENTRATION ABOUT MEDIAN
STATISTICAL MODEL COMPONENT
IDENTIFY DISTRIBUTIONAL MODEL
EXPONENTIAL
GAMMA
LOGNORMAL
NORMAL
WEIBULL
ESTIMATE PARAMETERS FROM DETERMINISTIC MODEL OUTPUT ABOUT MEDIAN
ESTIMATES OF ENTIRE DISTRIBUTION WITH CONFIDENCE INTERVALS

Figure 2.1: An example of the major steps in the application of the hybrid modelling methodology to area, line and point source emission regimes.

Figure 2.1 is intended to illustrate only how the hybrid modelling methodology might be applied. The deterministic and statistical models listed in Figure 2.1, while representing models actually developed in this thesis, are intended only as examples of the deterministic and statistical models which could be combined to form a hybrid model. However, the important steps in the hybrid model building process are indicated.

Of course some basis upon which the deterministic and statistical models can be selected is required. Unfortunately no theoretical analysis indicates which deterministic or statistical model is best for modelling air quality data. However, Pollack (1975) examined conditions which would lead to a lognormal distribution of air quality observations and Bencala and Seinfeld (1976) showed that for several possible distributional forms for windspeed data that the inverse of these distributions produced approximately lognormal distributions.

In view of the lack of theoretical guidance in the selection of models, an empirical basis for model selection was adopted. As noted earlier the simpler deterministic models were shown in many cases to provide a level of model performance equal to that of more complex models. As these models require fewer parameters and are more easily evaluated with a concomittant reduction in the use of computational resources, simpler models have been preferred. Of course the assumptions necessary to apply the models selected must be satisfied.

That models be commensurate with the amount and type of data available for analysis has also lead to the selection of simpler models. In Australia this is particularly important as only limited meteorological and emissions data are available. Thus the models developed here mostly rely on data collected on a routine basis. Of course when more detailed information becomes available more complex models can be constructed.

The selection of distributional models for inclusion within hybrid models can proceed in a more direct manner than that for the deterministic model component. Once a range of models which could be employed to describe the distribution of air quality data have been selected then the methods of statistical modelling may be applied (Gilchrist, 1984; Shapiro and Gross, 1981). The range of models selected represents those models commonly considered applicable to air quality data (Bencala and Seinfeld, 1976; Georgopoulos and Seinfeld, 1982). The models selected have been those with as few parameters as possible. In this way unnecessary parameterization has been avoided. The smallest number of parameters is particularly desirable as the parameters must be estimated from a restricted percentile range and then the upper percentiles of the distribution estimated by extrapolation. As the upper percentiles are determined by extrapolation, increasing the number of parameters will not necessarily reduce the uncertainty associated with predictions of these percentiles. This can arise because the reduced noise of the fit to the restricted percentile range resulting from the increased parameterization of the model appears as increased uncertainty in the parameter estimates.

The methods for the estimation of parameters for statistical distributions are presented in Chapter 3. The problems of identification of a distributional model in general and selection of a distributional model from amongst a range of alternatives are considered in detail in Chapter 4. The tests developed are independent of differences of scale allowing the results of a test on one data set to be readily compared with a test upon another. Following Chapter 4, it is demonstrated that the model selection procedure based upon goodness-of-fit tests has the desirable quality of selecting a model which produces good estimates of the upper percentiles of the distribution of air pollutant concentrations.

CHAPTER 3

STATISTICAL ESTIMATION OF THE PARAMETERS OF THE LOGNORMAL, GAMMA, WEIBULL AND EXPONENTIAL DISTRIBUTIONS

3.1 Introduction

When fitting distributions to observations of air quality data we are faced with several practical problems. We would ideally like to describe the entire distribution while determining the upper percentiles with greatest accuracy. Preferably we wish to characterize the data by only one or two parameter distributional forms. Such simple distributional forms may then be readily applied by those concerned with the assessment or management of air quality whose primary interest may be the frequency with which a pollutant concentration is equalled or exceeded. This information is usually required for a number of pollutants, such as carbon monoxide, nitrogen dioxide and ozone, and over a range of averaging times varying from several minutes to a full year. This information is usually required for a variety of source types including area, line and point sources.

In this chapter, methods are examined by which the parameters of the two-parameter lognormal, gamma and Weibull distributions, and the one parameter exponential distribution, may be estimated. Specific attention is given, for each method of parameter estimation, to the evaluation of the upper percentiles of each distribution. Also a simple method yielding an approximate confidence interval for the estimates of the upper percentiles of each distribution is derived. This approach does not require substantial numerical computation in order to provide an estimate of the confidence interval for a particular percentile.

3.2 Goodness-of-fit

In this chapter the problem of testing the hypothesis that the data are drawn from a particular distributional form is not considered. We are concerned with the effect that the method of parameter estimation has upon the evaluation of the percentiles of the distribution. The performance ~~criteria~~ criterion with which we shall assess the goodness-of-fit of the estimated distributions to the actual distributions is the root mean square error (rmse) evaluated at particular percentiles. Traditional statistical measures for assessing the suitability of a method of estimating the parameters of a distribution consider the bias or otherwise of the estimators and their variance about the true value of the parameter. It should be noted that such criteria are implicitly incorporated in the examination of the root mean square errors. The advantage, however, of examining the errors in the percentiles is that methods of parameter estimation which minimize the errors in fitting the percentiles in the region of greatest interest may be selected. Again for air quality data this will be the upper percentiles of the distribution. Clearly where management and policy decisions are based upon estimates of the upper percentiles then we wish to reduce losses in accuracy about these percentiles.

For the lognormal distribution the first two error or loss functions are the root mean square error (rmse) of estimation of the 1- and 99-percentiles, rmse $[\hat{x}_{0.01}]$ and rmse $[\hat{x}_{0.99}]$. These percentiles were chosen in order to evaluate the efficiency with which the more extreme events are described and also to provide an indication as to any bias associated with the estimators. The third loss function, designated L, is the average mean squared deviation between the actual and estimated distributions (Stedinger, 1980). This error function measures the overall deviation between the true and fitted distributions and could be considered an appropriate criteria for selecting the best estimators when the entire distribution of pollutant concentration is significant as may be the case where damages are evaluated. The error function is stated as follows (Stedinger, 1980)

$$L = E\left[\int_0^1 (x_p - \hat{x}_p)^2 dp\right] = \int_0^1 E[(x_p - \hat{x}_p)^2] dp \qquad (3.1)$$

where x_p is the true value of the p-percentile and $\hat{x}_p$ is the estimate of x_p. For the two-parameter lognormal distribution, x_p, is given by

$$x_p = \exp\ [a+bz_p] \qquad (3.2)$$

where z_p corresponds to the 100p percentile of the standard normal distribution and will henceforth be referred to as the standard variate. If $\hat{a}$ and $\hat{b}$ are the sample estimates of the location, a, and scale parameter, b, then

$$_0\int^1 (x_p - \hat{x}_p)^2 dp = \int_{-\infty}^{\infty} \{\exp [a + bz] - \exp(\hat{a} + \hat{b}z]\}^2 \frac{1}{(2\pi)^{1/2}} \exp [- \frac{z^2}{2}] \cancel{dt}\, dz \qquad (3.3)$$

$$_0\int^1 (x_p-\hat{x}_p)^2 dp = \exp [2a + 2b^2] + \exp [2\hat{a} + 2\hat{b}^2] - 2\exp [a + \hat{a} + \tfrac{1}{2}(b + \hat{b})^2] \qquad (3.4)$$

In order to estimate the effect of differences in magnitude of the root mean square error, relative errors were determined. These were chosen as the ratios L/σ_X^2 where σ_X^2 is the variance, rmse $[\hat{x}_{0.01}]/x_{0.01}$, and rmse$[\hat{x}_{0.99}]/ x_{0.99}$. Unbiased estimators should produce equivalent relative errors at all percentiles including the 1- and 99-percentiles.

As will be demonstrated the use of the error function L is of little value when examining the exponential distribution as the single parameter is estimated using two unbiased methods. For the gamma and Weibull distribution the bias of all methods of parameter estimation limits the usefulness of the error function L. Thus for the exponential, gamma and Weibull distributions three error criteria, rmse$[\hat{x}_{0.01}]/x_{0.01}$, rmse$[\hat{x}_{0.50}]/x_{0.50}$ and rmse$[\hat{x}_{0.99}]/x_{0.99}$ were evaluated.

3.3 Fitting the two-parameter lognormal distribution to air quality data

The lognormal distribution has been widely employed in the study of air pollution data over the past decade. While attracting a great deal of interest both as an empirical model and as a theoretical model only one study (Mage and Ott, 1984) has either reported findings on the errors in estimation of the percentiles associated with the application of the lognormal assumption, or has given attention to the methods by which the parameters of the distribution may be estimated. Of particular interest is the method attributed to Larsen (1971) where the upper percentiles of the distribution are fitted, assuming that the air pollutant data are adequately described as lognormal.

Reported here are the findings of a Monte Carlo study evaluating the behaviour of several methods of estimating the parameters in terms of the relative root mean square error at 1- and 99-percentiles and the average mean squared deviation between the true and fitted distributions. The study reports the minimum errors which may be expected for various estimators of the parameters when fitting data drawn from a lognormal distribution. The performance of the estimators are considered over a range of sample sizes equivalent to those that may be encountered when examining 24-h averaged data collected over one year. It will be shown that the results obtained may be readily extrapolated to larger sample sizes with empirical models developed using the results obtained for the smaller sample sizes.

3.3.1 Parameter estimation

Traditional statistical techniques for determining estimates of the location parameter, a, and scale parameter, b, are the method of moments and the method of maximum likelihood (Kendall and Stuart, 1973). The maximum likelihood estimates of a and b are given by

$$a = \bar{l} = \frac{1}{n} \sum_{i=1}^{n} l_i \tag{3.5}$$

and

$$b^2 = v_l^2 = \frac{1}{n} \sum_{i=1}^{n} (l_i - \bar{l})^2 \tag{3.6}$$

where $l_i = \ln(x_i)$. Thus the maximum likelihood estimators of a and b are equivalent to the moment estimates of the variate $l = \ln(x_i)$ using the biased estimate v_l^2 of the variance, σ_l^2, of l.

For the method of moments, the evaluation of

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (3.7)$$

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (3.8)$$

allows the determination of the estimators according to

$$\hat{a} = \ln\{\bar{x}/ [1+S_x^2/(\bar{x})^2]^{1/2}\} \qquad (3.9)$$

$$\hat{b}^2 = \ln [1 + S_x^2/ (\bar{x})^2] \qquad (3.10)$$

The method first applied by Larsen (1971) to the study of air pollutant data is derived from the ordinary least squares fit to the equation

$$l_p = M_g + S_g z_p \qquad (3.11)$$

where $l_p = \ln (x_p)$ are order statistics, M_g is the geometric mean, S_g the geometric standard deviation and z_p is the standard variate evaluated at the p-percentile. $\hat{a}$ is the intercept and $\hat{b}$ the slope of this linear equation and are given by

$$\hat{b} = S_g = \frac{n \sum_{i=1}^{n} z_i l_i - (\sum_{i=1}^{n} z_i)(\sum_{i=1}^{n} l_i)}{n \sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2} \qquad (3.12)$$

$$\hat{a} = M_g = \bar{l} - S_g\bar{z} \qquad (3.13)$$

Note that at $z=0$ equation (3.13) reduces to the maximum likelihood estimator of a and equation (3.12) similarly reduces to the equivalent maximum likelihood estimator of b, but only when using the full data set. However, given the different numerical techniques involved, it was considered worthwhile to assess the performance of these estimators as well. It should also be noted that such estimators assume implicitly that the distribution of pollutant concentrations is lognormal, that is, the data plots as a straight line on lognormal graph paper. If, for instance, data were drawn from the gamma distribution the curve would, in most cases, not be linear.

The primary advantage of such a method is that the estimators, $\hat{a}$ and $\hat{b}$, may be determined from a restricted set of percentiles. In particular such a method allows a lognormal model to be established for the upper percentiles of the distribution of frequency concentration. Larsen (1971) considered that the upper percentiles are well described by a lognormal model for several pollutants over a broad range of averaging times. Accordingly we have examined the performance of the estimators, given by equations (3.12) and (3.13) using a restricted range of percentiles. The upper 30% and 50% of available data were chosen. The Larsen (1971) method itself can be viewed as a special case of the least squares fit to certain percentiles. As initially developed, the method required two data points (usually the $p = 70.0$ and $p = 99.9$ percentiles) to estimate the parameters.

3.3.2 Monte Carlo simulation studies

The parameters of the lognormal distributions investigated are listed in Table 3.1. The range of parameter values was chosen to reflect those observed in studies of air quality data. Sample sizes of $n = 50$, 100, 200 and 365 were selected to represent samples drawn from a full year of 24-h average data. It will be shown later that the results obtained may be extrapolated to larger sample sizes. At each sample size the three methods of estimation were examined. The estimators are $(\bar{x}, S_x^2)$, $(\bar{l}, V_l^2)$ and $(\ln M_g, \ln S_g)$. Table 3.2 presents the values of the ratios L/σ_x^2, rmse $[x_{0.01}]/x_{0.01}$ and rmse $[x_{0.99}]/x_{0.99}$ for each of the estimators at each sample size for the lognormal distributions considered. The results of Table 3.2 were obtained from 5000 Monte Carlo experiments. 5000 Monte Carlo experiments were undertaken in order to

Table 3.1: Parameters and percentiles of the lognormal distributions used in the Monte Carlo experiments

a	b	$x_{0.01}$	$x_{0.99}$	Coefficient of variation	Coefficient of skewness
3.00	0.1	15.917	25.346	0.100	0.310
3.00	0.3	9.995	40.363	0.307	1.028
3.00	0.5	6.277	64.275	0.533	1.945
3.00	0.7	3.941	102.355	0.795	3.191

obtain estimates of the rmse values to a precision of about three significant figures.

Examining Table 3.2 we note in general for all estimators that the error measurements L/σ_x^2, rmse $[\hat{x}_{0.01}]/x_{0.01}$ and rmse $[\hat{x}_{0.99}]/x_{0.99}$ increase with the value of b while decreasing with increased sample size. For all sample sizes n>50 the estimators $(\bar{1}, V_1^2)$ and $(\ln M_g, \ln S_g)$ are almost equivalent at the 1- and 99-percentiles and the root mean square criteria indicates that $(\bar{1}, V_1^2)$ and $(\ln M_g, \ln S_g)$ provide the best estimators of the parameters of the underlying distribution. The results of Table 3.2 for n=365 indicate that even for very large sample sizes n>365 the use of the estimators $(\bar{x}, S_x^2)$ will produce significantly larger relative errors than the estimators $(\bar{1}, V_1^2)$ and $(\ln M_g, \ln S_g)$.

So the estimators $(\bar{1}, V_1^2)$ and $(\ln M_g, \ln S_g)$ produce essentially equivalent results for large sample sizes. The estimators also appear to be unbiased as indicated by the equivalent magnitude of the errors at the $x_{0.01}$ and $x_{0.99}$ percentiles which is to be expected from theoretical considerations (Kendall and Stuart, 1973). Given then the relatively low computational effort required to obtain the estimates $(\bar{1}, V_1^2)$ these estimators would be preferred.

The estimators $(\ln M_g, \ln S_g)$ have also been investigated where the estimators are derived from the upper 30 and 50 percentiles of the distributions. The parameters of the lognormal distributions used in the analysis are again those of Table 3.1. The error criteria are as

TABLE 3.2: Results of fitting the lognormal distribution by three methods. The numbers are the result of Monte Carlo calculations with 5000 samples

	$L/\sigma_x{}^2$				rmse $[\hat{x}_{0.01}]/x_{0.01}$				rmse $[\hat{x}_{0.99}]/x_{0.99}$			
	b				b				b			
Estimators used	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
Sample Size n = 50												
$\bar{x}$, S_x^2	12.76	17.68	33.64	82.4	0.0280	0.091	0.175	0.297	0.0276	0.088	0.160	0.241
$\bar{l}$, V_l^2	12.68	16.37	26.93	55.1	0.0279	0.085	0.144	0.206	0.0273	0.082	0.137	0.194
ln (M_g), ln (S_g)	12.71	16.31	26.67	54.1	0.0279	0.084	0.143	0.204	0.0274	0.082	0.136	0.192
Sample Size n = 100												
$\bar{x}$, S_x^2	6.10	8.63	16.96	43.8	0.0192	0.064	0.127	0.220	0.0193	0.062	0.115	0.179
$\bar{l}$, V_l^2	6.05	7.86	12.94	26.2	0.0191	0.058	0.097	0.138	0.0191	0.057	0.095	0.134
ln (M_g), ln (S_g)	6.06	7.85	12.88	25.9	0.0191	0.058	0.097	0.137	0.0191	0.057	0.095	0.133
Sample Size n = 200												
$\bar{x}$, S_x^2	3.21	4.49	8.85	23.83	0.0141	0.047	0.094	0.168	0.0138	0.044	0.083	0.132
$\bar{l}$, V_l^2	3.18	4.09	6.65	13.30	0.0140	0.042	0.070	0.099	0.0136	0.041	0.068	0.095
ln (M_g), ln (S_g)	3.19	4.09	6.64	13.24	0.0140	0.042	0.070	0.099	0.0136	0.041	0.068	0.095
Sample Size n = 365												
$\bar{x}$, S_x^2	1.70	2.47	4.85	13.83	0.0102	0.034	0.071	0.129	0.0101	0.033	0.062	0.102
$\bar{l}$, V_l^2	1.69	2.24	3.58	7.29	0.0101	0.030	0.050	0.070	0.0100	0.030	0.050	0.070
ln (M_g), ln (S_g)	1.69	2.23	3.57	7.26	0.0101	0.030	0.050	0.070	0.0100	0.030	0.050	0.070

stated previously and were evaluated for sample sizes of n=50, 100, 200 and 365. Again for all distributions at each sample size, 5000 Monte Carlo experiments were performed from which the error criteria were evaluated. The results are given in Table 3.3. It should be noted that all three measurements of error increase with the increasing magnitude of the scale parameter and decrease with larger sample size.

Comparing values of the rmse $[\hat{x}_{0.99}]/x_{0.99}$ criterion where the estimators are derived for the upper 50% of the data with that obtained when fitting all the data (see Table 3.2) we find that the value for the upper 50% estimators is almost equivalent to that found when fitting all the data when this is undertaken at half the sample size. This is not, however, the case for rmse $[\hat{x}_{0.01}]/x_{0.01}$ where substantial increases in error are recorded as might be expected given that the upper percentiles form the basis of the parameter estimation.

When estimating the parameters of the distribution using only the upper 30% of the data increases in all error criteria over that obtained for the upper 50% of the data resulted. These increases are particularly marked for the overall measure of error, L/σ_x^2, and rmse $[\hat{x}_{0.01}]/x_{0.01}$ reflecting the inherent bias of the estimation procedure.

3.3.3 The relative root mean square errors associated with the estimation of the percentiles of the lognormal distribution

The results of Tables 3.2 and 3.3 represent the minimum root mean square errors that may be expected for 1- and 99-percentiles when fitting the two-parameters of the distribution to data where the parameters of the distribution are unknown. An empirical model of these errors for $x_{0.99}$ using the maximum likelihood method results for varying scale parameter σ, and sample size, n, has been derived. The model allows estimates of this error to be derived for sample sizes and sample scale parameters other than those listed in Table 3.2. This estimate of the error is the minimum to be expected when used in conjunction with air quality data where parameters are evaluated from the sample and the assumption of lognormality may hold only approximately.

Table 3.3: Results based on the estimators ($\ln M_g$, $\ln S_g$) derived using the upper percentiles of the available data.

b	L/σ_X^2		rmse $[\hat{x}_{0.01}]/x_{0.01}$		rmse $[\hat{x}_{0.99}]/x_{0.99}$	
	50%	30%(a)	50%	30%	50%	30%
			Sample Size n = 50			
0.1	25.6	51.2	0.054	0.082	0.034	0.038
0.3	28.0	50.9	0.169	0.267	0.102	0.114
0.5	40.6	64.0	0.298	0.475	0.168	0.190
0.7	83.5	118.8	0.442	0.770	0.239	0.263
			Sample Size n = 100			
0.1	13.0	26.3	0.038	0.059	0.024	0.027
0.3	13.8	25.1	0.117	0.184	0.072	0.080
0.5	20.4	31.4	0.202	0.320	0.120	0.133
0.7	41.0	56.6	0.293	0.489	0.168	0.185
			Sample Size n = 200			
0.1	6.4	13.3	0.027	0.042	0.017	0.019
0.3	6.9	12.5	0.083	0.127	0.052	0.057
0.5	10.4	15.8	0.142	0.219	0.085	0.095
0.7	20.8	27.9	0.201	0.323	0.120	0.131
			Sample Size n = 365			
0.1	3.5	7.3	0.020	0.031	0.012	0.014
0.3	3.8	6.8	0.060	0.094	0.038	0.042
0.5	5.6	8.7	0.104	0.160	0.062	0.071
0.7	11.5	15.6	0.148	0.230	0.089	0.099

(a) Upper percentiles used to derive estimators of the scale and location parameters of the distribution.

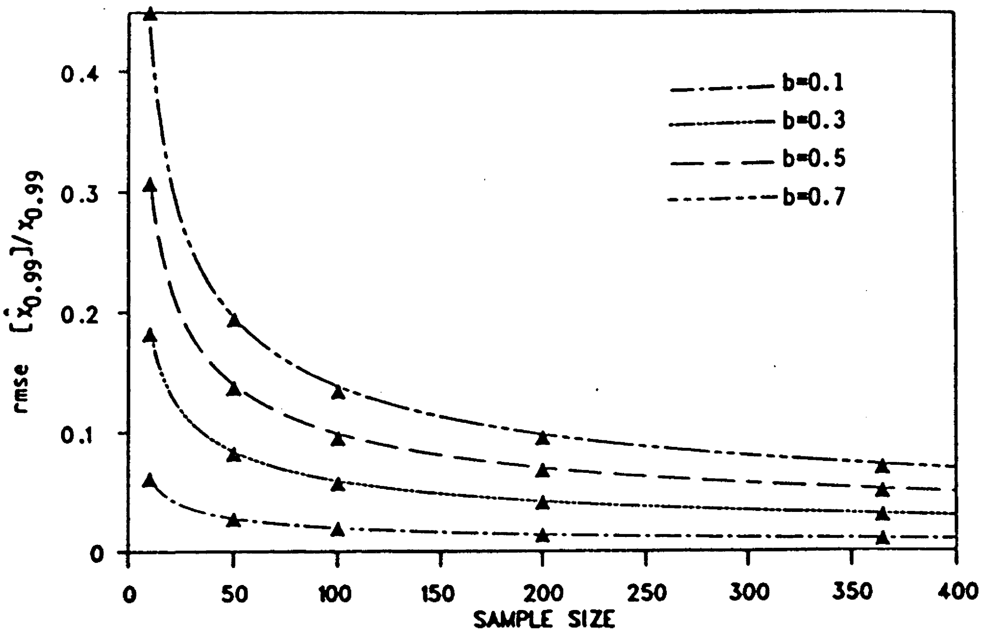

Figure 3.1: The fit according to the empirical model (equation 3.14) to the rmse $[x_{0.99}]/x_{0.99}$ data of Table 3.2 for the maximum likelihood estimators of the lognormal distribution parameters.

The basic empirical model is

$$F(\sigma,n) = \frac{\kappa\,\sigma}{\sqrt{n}} \qquad (3.14)$$

where $F(\sigma,n)$ in this case is rmse $[\hat{x}_{0.99}]/x_{0.99}$, σ is the scale parameter of the underlying lognormal distribution, n is the sample size and κ is a constant.

The value of κ may vary with the percentile of the distribution under consideration. The constant κ was determined by non-linear least squares methods using an algorithm based on a golden section search. For the 99-percentile a value of $\kappa = 1.975$ was calculated. Figure 3.1 illustrates the fit of the model (3.14) to the rmse $[\hat{x}_{0.99}]/x_{0.99}$ data of Table 3.2 for the estimators $(\bar{1},V_1^2)$. It should be noted that equation (3.14) also provides an excellent model with which to approximate rmse $[\hat{x}_{0.01}]/x_{0.01}$ for the estimators $(\bar{1}, V_1^2)$ and for the rmse errors associated with the estimators $(\ln M_g, \ln S_g)$ at larger sample sizes.

3.3.4 Conclusions for the lognormal distribution

When fitting lognormal distributions to air quality data we note for the range of sample sizes considered that the maximum likelihood estimators $(\bar{l}, V_l^2)$ and the estimators $(\ln M_g, \ln S_g)$ provide the best estimators of the upper percentiles of the lognormal distribution. Where the data are considered to be drawn from a lognormal distribution fitting all the data will yield the best overall estimates of the distribution including the upper percentiles.

However for larger sample sizes relatively small increases in rmse $[\hat{x}_{0.99}]/x_{0.99}$ occur. A convenient empirical model was found for determining an approximation to the error associated with the estimation of a particular percentile where the estimators of the parameters of the lognormal distribution are $(\bar{l}, V_l^2)$ and $(\ln M_g, \ln S_g)$. Where the parameters of the distribution must be estimated from the sample such a model may still yield a useful estimate of the possible error. When considering air quality data which has been tested for goodness-of-fit to the lognormal form then this estimate might be adopted as the likely minimum error. It should also be noted that the empirical model developed here for the 1- and 99-percentiles may be constructed at other percentiles of interest based on suitable sampling studies.

3.4 Fitting the exponential distribution to air quality data

The exponential distribution has been of interest in the study of air quality data partly due to its application to describe pollutant data observed about isolated point sources (Curran and Frank, 1975; Simpson, Butt and Jakeman, 1984). The two-parameter exponential model (Berger et al., 1982) has also been applied to observations of pollutant concentrations where the source is a combination of several point sources in conjunction with an area source. While the exponential distribution has not been established as a model applicable to a wide range of pollutants at various averaging times, there exists sufficient interest in this model to warrant examination of the estimation of the parameter of this distribution and the variance associated with the estimate of particular percentiles over a range of sample sizes.

3.4.1 Parameter estimation

The exponential distribution has the scale parameter, b, which must be estimated. The usual method of estimation is to use the arithmetic mean, which is the maximum likelihood estimator of this parameter and is stated thus,

$$\hat{b} = \bar{x} = \sum_{i=1}^{n} x_i \qquad (3.15)$$

Another useful estimate of the scale parameter can be derived from the relationship between the median, M_a, and mean b of an exponential distribution where

$$b = M_a \,/\, \ln 2 \qquad (3.16)$$

so that the estimator is

$$\hat{b} = 1.4427\ M_a \qquad (3.17)$$

This method of estimation will of course be more useful if the estimate of the median is more accurate than the mean, for example when the sample distribution contains outliers.

3.4.2 Monte Carlo simulation studies

Since the relative root mean square errors are evaluated the results obtained will be independent of the value of the scale parameter. Thus for all Monte Carlo experiments the underlying scale parameter of the exponential distribution was chosen as b=1. Sample sizes of n=50, 100, 200 and 365 were selected to represent a range of sample sizes which might be observed in a year of monitoring 24-h average pollutant levels. It will be shown that the results obtained here may be readily extrapolated to larger sample sizes. At each sample size the two methods of estimation of b, as given by equations (3.15) and (3.17) , were

Table 3.4: Results of estimation of the scale parameter of the exponential distribution based on 5000 Monte Carlo experiments.

Sample size	rmse$[\hat{x}_p]/x_p$	
	$\bar{x}$	1.4427 M_a
50	0.1410	0.2045
100	0.0995	0.1418
200	0.0692	0.1015
365	0.0518	0.0749

examined using 5000 Monte Carlo experiments. These results appear in Table 3.4. It should be noted that for rmse$[\hat{x}_p]/x_p$ at p = 1, 50 and 99 equivalent results for both methods of parameter estimation were obtained.

The results of Table 3.4 indicate, as would be expected, that in all cases both estimators are unbiased. The estimator based on the sample mean is clearly that with the lowest variance. The estimator based on the median has produced rmse results which are not substantially greater than those based on the sample mean. These results illustrate the possible usefulness of the median as a basis for the estimation of the parameter of the exponential distribution where either the sample mean may not be calculated directly or the sample distribution contains outliers.

3.4.3 The relative root mean square errors associated with the estimation of thẹ percentiles of the exponential distribution

A useful empirical model for determining the relative rmse at all percentiles and for sample sizes n>49 was also developed for the exponential distribution. The model form identified is:-

$$F(n) = \frac{\kappa}{\sqrt{n}} \qquad (3.18)$$

where F(n) is $\mathrm{rmse}[\chi_p]/\chi_p$ at the p-percentile and n is the sample size. The constant κ was calculated as the average value of the four values of $F(n)\sqrt{n}$ derived from Table 3.4 for both estimators. These values are for the estimator based on the sample mean, $\kappa = 0.990$, and for the estimator based on the sample median $\kappa = 1.433$. The results of Table 3.4 with the fit according to equation (3.18) with the appropriate value of constant κ are presented as Figure 3.2.

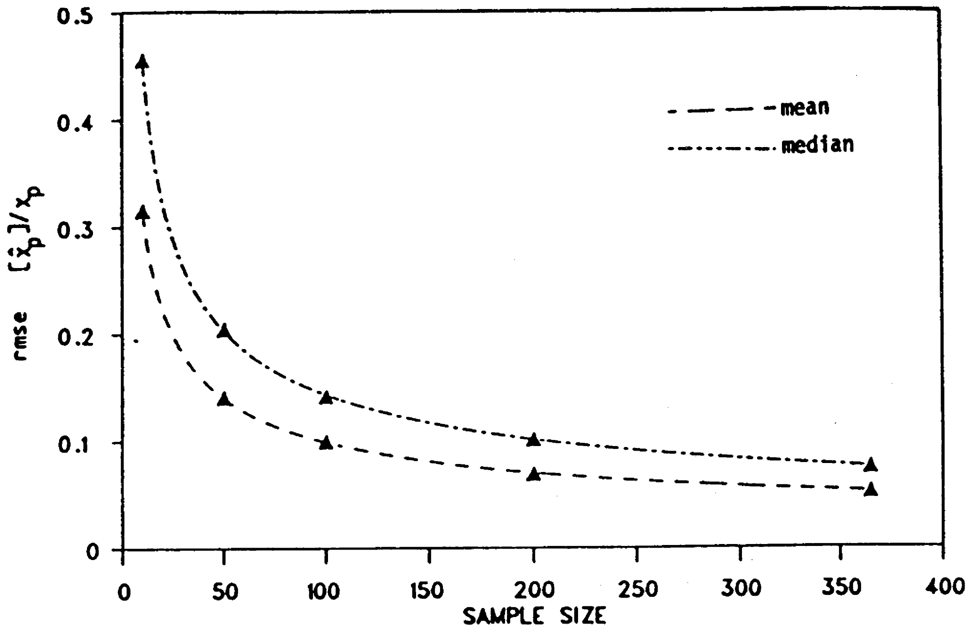

Figure 3.2: The fit according to the empirical model (equation 3.18) to the $\mathrm{rmse}\ [\hat{\chi}_p]/\chi_p$ data of Table 3.4 for the estimators of the exponential distribution parameter based upon the mean and the median.

3.4.4 Conclusions for the exponential distribution

Where the full sample distribution is available and reliable, the sample mean provides the most accurate estimate of the exponential distribution parameter. However, when modelling the distribution of pollutant concentrations the median concentration may provide a more accurate estimate of the scale parameter. Therefore where an exponential model is considered applicable only to a limited range of percentiles including the median or the data contains outliers, then the median may provide a useful estimate of the scale parameter for such a model. For both estimators equation (3.18) with the appropriate value of κ should yield a useful estimate of the minimum variance to be expected when estimating a percentile of the distribution.

3.5. Fitting the two-parameter gamma distribution to air quality data

The gamma distribution has found application in air pollution studies as the distribution is skewed to the right, as is the case for most air quality data. The gamma distribution has generally been considered to be of importance in the study of the statistical distribution of air quality data (Bencala and Seinfeld, 1976; Pollack, 1975). Berger et al. (1982) found that a gamma distribution may provide a better description of 24-h average sulphur dioxide concentrations than the lognormal distribution.

While the gamma distribution has been found to be of value as a useful distributional model no studies have reported the errors associated with the estimation of the percentiles. In fact, not even the estimation of the parameters of the distribution has been examined from the point of view of air quality studies. In this study three methods of estimation of the parameters of the distribution are examined and their performance ~~are~~ is evaluated at the 1-, 50- and 99-percentiles using the relative root mean square error (rmse) as the performance criteria. These performance criteria were evaluated over a range of sample sizes which might represent those encountered when examining 24-h averaged data. However the results obtained may be readily extrapolated to larger sample sizes using a simple empirical model.

Known methods of parameter estimation for the gamma distribution examined in this study are the method of maximum likelihood, the method of moments and a modified method of moments. These methods have been examined in the study of flood frequency analysis (Nozdryn-Plotnicki and Watt, 1979) but the form of the distribution widely used in hydrology is not of great interest in air quality studies as it is the three parameter form which is applied in conjunction with a logarithmic transformation of the hydrologic data. A method proposed by Bobee (1975) applicable to such distributions where the moments of the untransformed data are preserved was found by Nozdryn-Plotnicki and Watt (1979) to yield best results only where the scale parameter is negative in the form of gamma distribution examined. They considered both the method of moments and the method of maximum likelihood as superior procedures for the estimation of the parameters of the gamma distribution.

The efficiency with which the shape parameter of the gamma distribution may be estimated by the method of moments relative to the maximum likelihood method was found to be as low as 22% (Kendall and Stuart, 1973). Efficiency is the ratio of the variance of the estimates of the parameters derived from two methods (in this case the method of moments and the method of maximum likelihood). While this result indicates that the method of maximum likelihood should be favoured as the method of estimation the ability of both methods to estimate the percentiles of the distribution is examined. In particular the concern is with the upper percentiles of the distribution.

Finally it should be noted that the maximum likelihood estimators produce biased estimates of the parameters of the gamma distribution. This result was inferred from numerical studies by Choi and Wette (1969) who considered that the estimators were positively biased. Shenton and Bowman (1972) obtained the asymptotic expansions for the biases of the parameters. Anderson and Ray (1975) extended the work of Shenton and Bowman (1972) and obtained modified maximum likelihood estimators to yield approximately unbiased estimators. Finally Berman (1981) showed from theoretical considerations that the method of maximum likelihood yields estimators that are always positively biased. It should be noted that the bias of the maximum likelihood estimators decreases with increasing sample size.

3.5.1 Parameter estimation for the gamma distribution

The probability density function of the random variable denoting pollutant concentration, χ, described as having a gamma distribution with parameters α and β , termed the scale and shape parameters respectively is

$$f(\chi) = (\chi/\alpha)^{\beta-1}[\exp(-\chi/\alpha)]/\alpha\Gamma(\beta) \qquad (3.19)$$

where Γ denotes the gamma function.

Let $\chi_1,\chi_2,\ldots,\chi_n(n>1)$ represent a random sample of values of χ and let L denote the log likelihood function. Then we have

$$L = n[\beta\log\alpha - \log\Gamma(\beta)] + (\beta-1)\sum_{i=1}^{n}\log\chi_i - \alpha\sum_{i=1}^{n}\chi_i \qquad (3.20)$$

and taking the partial derivatives with respect to the parameters α and β we obtain

$$\frac{\partial L}{\partial\beta} = n\,[\log\alpha - \Psi\,(\beta)] + \sum_{i=1}^{n}\log\chi_i \qquad (3.21)$$

$$\frac{\partial L}{\partial\alpha} = n\,(\beta/\alpha) - \sum_{i=1}^{n}\chi_i \qquad (3.22)$$

where

$$\Psi\,(\beta) = \frac{d}{d\beta}\log\Gamma(\beta) \qquad (3.23)$$

and is referred to as the digamma function. Analytical solutions for the maximum likelihood estimators cannot be derived and thus the solution of

$\partial L/\partial\alpha = 0$ and $\partial L/\partial\beta = 0$ yielding the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ must be obtained numerically. The method selected was the Newton-Raphson method (Choi and Wette, 1969). The solution of equations (3.21) and (3.22) for $\hat{\beta}$ produces the equation

$$\log \hat{\beta} - \Psi(\hat{\beta}) = M \qquad (3.24)$$

where

$$M = \sum_{i=1}^{n} X_i/_{n} - (\sum_{i=1}^{n} \log X_i)/n \qquad (3.25)$$

The Newton-Raphson iteration method then gives

$$\hat{\beta}_k = \hat{\beta}_{k-1} - \frac{\log \hat{\beta}_{k-1} - \Psi(\hat{\beta}_{k-1}) - M}{1/\hat{\beta}_{k-1} - \psi^{-1}(\hat{\beta}_{k-1})} \qquad (3.26)$$

where $\hat{\beta}_k$ denotes the kth estimate starting with the initial value $\hat{\beta}_0$ and $\Psi'(\beta)$ represents $d\Psi(\beta)/d\beta$ which is referred to as the trigamma function. Jordan (1960) gives good approximations to the digamma and trigamma functions as

$$\Psi(\beta) \sim \log \beta - \{1 + [1- (0.1 - 1/(21))/\beta^2]/6\beta \}/(2\beta) \qquad (3.27)$$

and

$$\Psi'(\beta) \sim \{1 + \{ 1 + [1 - (0.2 - 1/(7\beta^{2}))/\beta^2]/(3\beta)\}/(2\beta)\}/\beta \qquad (3.28)$$

Once $\hat{\beta}$ is determined, α may be estimated from

$$\hat{\alpha} = \hat{\beta} / \sum_{i=1}^{n} x_i / n \qquad (3.29)$$

The method of moments is based on the assumption that the sample moments should provide good estimates of the corresponding population moments. The estimator of the gamma scale parameter by the method of moments, here designated $\hat{a}$, is given by

$$\hat{a} = S^2 / \bar{x} \qquad (3.30)$$

where $\bar{x}$ is the sample mean and S^2 is the unadjusted sample variance. The shape parameter of the gamma distribution is, by the method of moments, estimated by

$$\hat{b} = (\bar{x}/S)^2 \qquad (3.31)$$

It should be noted that the unadjusted sample variance is given by

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (3.32)$$

In addition to the maximum likelihood and method of moments techniques it was also considered worthwhile to examine the application of the unbiased estimates of the sample variance to the calculation of the estimators of the gamma distribution. The unbiased estimator of the sample variance is given by

$$S_u^{\,2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (3.33)$$

so the estimator of the gamma scale parameter, $\hat{a}_u$, is given by

$$\hat{a}_u = S_u^2/\bar{x} \qquad (3.34)$$

and the shape parameter, $\hat{b}_u$, by the equation

$$\hat{b}_u = (\bar{x} \ / \ S_u)^2 \qquad (3.35)$$

3.5.2 Monte Carlo simulation studies

The parameters of the gamma distributions examined in this thesis are given in Table 3.5. The range of parameter values selected reflects that which may be observed from air quality data. Sample sizes of n=50, 100, 200, 365 were selected to represent the range of sample sizes which may be obtained in a year of recording 24-h average data. It should be noted that the results obtained may be extrapolated to larger sample sizes. At each sample size considered the three methods of estimation were examined. The pairs of estimators are $(\hat{\alpha},\hat{\beta})$, $(\hat{a},\hat{b})$ and $(\hat{a}_u,\hat{b}_u)$. Table 3.6 presents the values of the relative root mean square error $[\hat{x}_p]/x_p$ at the p= 1-, 50- and 99-percentiles for each of the estimators at each sample size for all the gamma distributions considered. The results listed in Table 3.6 were obtained as the result of 5000 Monte Carlo experiments.

Examining Table 3.6 it should be noted that the estimators $(\hat{a}_u, \hat{b}_u)$ based upon the unbiased sample variance provide improved estimates of the percentiles of the distributions in all cases over that of the method of moments, $(\hat{a}, \hat{b})$. For all sample sizes the method of maximum likelihood provides the superior estimates of the parameters of the gamma distribution as reflected in the reduced root mean square errors at all percentiles and over the range of the shape parameter considered. In particular the maximum likelihood estimators provide significantly improved estimates of the 1- and 50-percentiles.

Table 3.5: Parameters and percentiles of the gamma distributions used in the Monte Carlo experiments.

α	β	$x_{0.01}$	$x_{0.99}$	Coefficient of variation	Coefficient of skewness
1.00	1.00	0.010	4.605	1.000	2.000
1.00	2.00	0.149	6.638	0.707	1.414
1.00	3.00	0.436	8.406	0.577	1.155
1.00	4.00	0.823	10.045	0.500	1.000

The positive bias of all three methods of estimation of the parameters of gamma distribution may be clearly observed in Table 3.6. Here the relative root mean square errors decrease as we move successively from the 1- to the 50- and 99-percentiles. It appears from the small change in $[\hat{x}_{0.99}]/x_{0.99}$ with the changing shape parameter of the gamma distribution that the positive bias does not affect as significantly the estimation of the upper percentiles of the distribution.

3.5.3 The relative root mean square errors associated with the estimation of the percentiles of the gamma distribution

The results listed in Table 3.6 represent the minimum relative root mean square errors that may be expected when fitting the two-parameter gamma distribution where the parameters of the distribution are unknown. For the 50- and 99-percentiles an empirical model has been developed to provide an estimate of $\text{rmse}\ [\hat{x}_p]/x_p$ based upon the following general expression

Table 3.6: Results of estimating the parameters of the gamma distribution by three methods. The numbers are the results of 5000 Monte Carlo experiments.

Estimators	$\lvert\hat{X}_{0.01}\rvert/X_{0.01}$				$\lvert\hat{X}_{0.50}\rvert/X_{0.50}$				$\lvert\hat{X}_{0.99}\rvert X_{0.99}$			
	Shape Parameter				Shape Parameter				Shape Parameter			
	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000
						Sample Size n=50						
$(\hat{a}_u,\hat{b}_u)$	2.793	0.930	0.665	0.561	0.411	0.306	0.282	0.274	0.139	0.141	0.148	0.157
$(\hat{a},\hat{b})$	3.021	0.993	0.705	0.591	0.430	0.319	0.294	0.285	0.144	0.146	0.154	0.162
$(\hat{\alpha},\hat{\beta})$	1.699	0.737	0.591	0.525	0.279	0.245	0.250	0.253	0.095	0.113	0.131	0.144
						Sample Size n=100						
$(\hat{a}_u,\hat{b}_u)$	1.384	0.581	0.425	0.363	0.2771	0.206	0.189	0.184	0.097	0.096	0.100	0.106
$(\hat{a},\hat{b})$	1.455	0.603	0.439	0.374	0.283	0.211	0.193	0.188	0.099	0.098	0.102	0.108
$(\hat{\alpha},\hat{\beta})$	0.889	0.439	0.373	0.326	0.184	0.160	0.166	0.165	0.064	0.075	0.088	0.095
						Sample Size n=200						
$(\hat{a}_u,\hat{b}_u)$	0.879	0.385	0.289	0.248	0.203	0.144	0.132	0.128	0.072	0.068	0.072	0.075
$(\hat{a},\hat{b})$	0.904	0.393	0.294	0.252	0.205	0.145	0.134	0.130	0.072	0.068	0.072	0.075
$(\hat{\alpha},\hat{\beta})$	0.507	0.298	0.252	0.217	0.125	0.113	0.115	0.112	0.044	0.053	0.062	0.065
						Sample Size n=365						
$(\hat{a}_u,\hat{b}_u)$	0.601	0.273	0.206	0.171	0.151	0.105	0.095	0.089	0.054	0.049	0.051	0.052
$(\hat{a},\hat{b})$	0.611	0.276	0.208	0.173	0.152	0.105	0.096	0.090	0.054	0.050	0.052	0.052
$(\hat{\alpha},\hat{\beta})$	0.371	0.211	0.175	0.150	0.096	0.081	0.081	0.078	0.034	0.038	0.043	0.046

$$F(n) = \frac{\kappa}{\sqrt{n}} \qquad (3.36)$$

where $F(n)$ is $\mathrm{rmse}[\hat{x}_p]/x_p$, n is the sample size and $\kappa = 1.087$ was determined as the mean value of $F(n)\sqrt{n}$ for the data available in Table 3.6.

Expressions based on equation (3.36) were developed with the values of $\mathrm{rmse}[\hat{x}_p]/x_p$ obtained using the method of maximum likelihood. The empirical model was developed for sample sizes of n>10, though the model provides the best approximations for sample sizes of $n > 50$. Figure 3.3 presents the fit of the empirical model, equation (3.36), to the $\mathrm{rmse}[\hat{x}_{0.99}]/x_{0.99}$ of Table 3.6. It should be noted that the empirical model does not account for the small change in $\mathrm{rmse}\,[x_{0.99}]/x_{0.99}$ that occurs with the change in the gamma shape parameter. However equation (3.36) should provide a useful approximation for air quality data as this small error may, for most applications, not be significant.

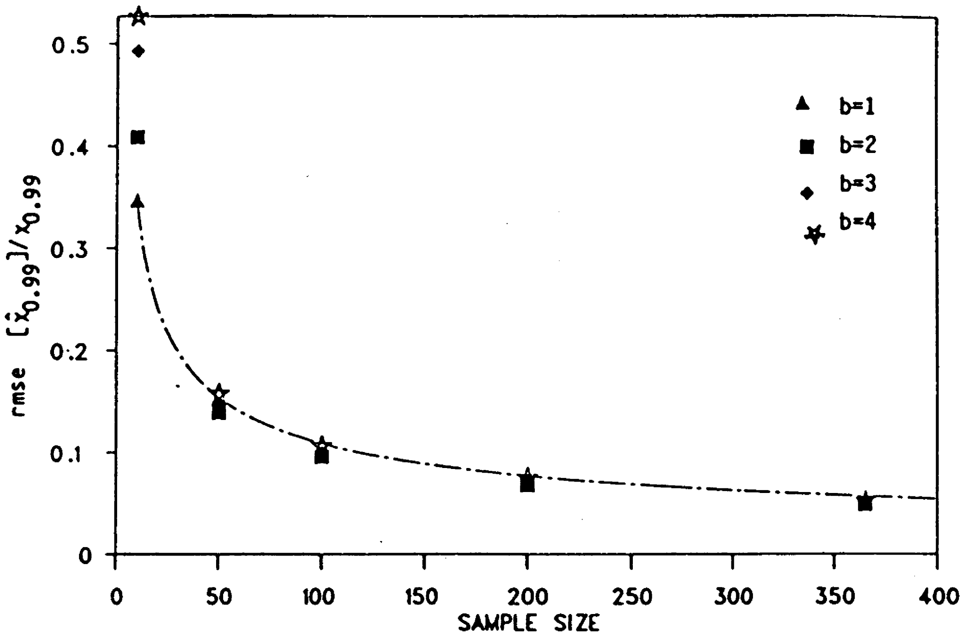

Figure 3.3: The fit according to the empirical model (equation 3.36) to the $\mathrm{rmse}\,[\hat{x}_{0.99}]/x_{0.99}$ data of Table 3.6 for the maximum likelihood estimators of the gamma distribution parameters.

3.5.4 Conclusions for the gamma distribution

The method of maximum likelihood provides the best estimators of the parameters of the gamma distribution. For large sample sizes, the method of moments may provide a more convenient approach to the estimation of the parameters of the gamma distribution. This is proposed as the estimators based on the method of moments can be rapidly computed with little loss of accuracy over that of the method of maximum likelihood. This is particularly the case when considering the estimates of the upper percentiles of the distribution.

It should be noted that where the method of moments is applied the unbiased estimators $(\hat{a}_u, \hat{b}_u)$ should be used in favour of the traditional estimators $(\hat{a}, \hat{b})$. In all cases the $(\hat{a}_u, \hat{b}_u)$ will provide improved estimates of the percentiles of the distribution.

The positive bias of the maximum likelihood estimators appears to predominantly affect estimation of the lower percentiles of the distribution. The application of unbiasing factors does not therefore appear warranted where the concern is with the upper percentiles of the distribution. Finally it should be noted that the effect of bias will be less significant for large sample sizes.

3.6 Fitting the two-parameter Weibull distribution to air quality data

Like the gamma distribution the Weibull distribution (Weibull, 1951) has found application in air quality studies because of its similarity in form to the lognormal distribution in that both distributions are positively skewed. The Weibull distribution takes the exponential form when the shape parameter has the value 1. Bencala and Seinfeld (1976) examined the sum of squares error in fitting the Weibull distribution to carbon monoxide data from 8 cities. Using this criteria the Weibull model produced lower values than that obtained by Larsen (1971) when fitting the two-parameter lognormal distribution for five of the eight data sets. In Chapter 6 the Weibull distribution is applied to model carbon monoxide dispersed from roadway line sources. Pollack (1975) also considered the Weibull distribution as relevant to air quality studies, though predominantly due to its similarity to the lognormal distribution. Pollack (1975) did not present evidence from the examination of air quality data as to the validity of the application of the Weibull distribution.

This section reports a study of the estimation of the parameters of the Weibull distribution by the method of maximum likelihood, the method of moments, a method attributed to Menon (1963) and an ordinary least squares method. In particular we examine the errors associated with estimating the percentiles of the distribution where the parameters must be estimated from the sample.

3.6.1 Parameter estimation

Defining the Weibull distribution function with two parameters, σ the scale parameter and p the shape parameter as

$$F\ (\chi) = 1 - \exp\ [-(\chi/\sigma)^{p}] \qquad (3.37)$$

Letting $x_1, x_2, \ldots, x_n$ be a random sample from the two-parameter Weibull distribution with parameters σ and p then Thoman et al. (1969) demonstrated that the maximum likelihood estimators of σ and p are obtained by solving the equations

$$\frac{n}{\hat{p}} - \frac{n\Sigma x_i^{\hat{p}}\ \ln x_i}{\Sigma x_i^{\hat{p}}} + \Sigma\ \ln x_i = 0 \qquad (3.38)$$

and

$$\hat{\sigma} = (\Sigma\ x_i^{\hat{p}}/n)^{1/\hat{p}} \qquad (3.39)$$

Equation (3.38) can be solved using the Newton-Raphson iterative method to yield the estimate $\hat{p}$ which on substitution into equation (3.39) yields the $\hat{\sigma}$ estimator.

The method of moments estimators, here designated as p_m and σ_m, are derived from the solution of the equation giving the coefficient of variation, V_p, as

$$V_p = \left[\frac{\Gamma((p_m+2)/p_m) - \Gamma^2((p_m+1)/p_m)}{\Gamma^2((p_m+1)/p_m)}\right]^{1/2} \tag{3.40}$$

where $V_p = s/\bar{x}$, with $\bar{x}$ being the sample mean and s^2 the variance. Equation (3.40) may be solved to yield p_m using an iterative technique. We then derive the moment estimator of σ from

$$\sigma_m = \left[\frac{\bar{x}}{\Gamma((p_m+1)/p_m}\right]^{p_m} \tag{3.41}$$

It should be noted that tables of p_m corresponding to V_p can be constructed using equation (3.40). Sinha and Kale (1980) provide tables of V_p for p_m = 0.1(0.05)4.00 which using interpolation would provide moment estimates with an accuracy of two decimal places. Such a procedure might be adopted where access to computing facilities is limited.

Menon (1963) obtained the method of moments estimates for σ and p, here denoted $\hat{\sigma}_e$and $\hat{p}_e$, using the distribution of the log-Weibull variable. These estimators are

$$\hat{p}_e = [(6/\pi^2)\ S_y{}^2]^{-1/2} \tag{3.42}$$

and

$$\hat{\sigma}_e = \exp\ [\bar{y} + 0.5772\ (1/\hat{p}_e)] \tag{3.43}$$

where $y_i = \log\ x_i$, i=1, ..., n and $\bar{y}$ and S_y are the usual sample mean and standard deviation.

Using a transformation of the Weibull distribution function, equation (3.37), the following equation results

$$\log \chi = \log \sigma + (1/p) \log [-\log(1-F(\chi))] \qquad (3.44)$$

which may readily be written in linear form as

$$y = \alpha + \beta x \qquad (3.45)$$

where $y = \log \chi$, $\alpha = \log \sigma$, $\beta = 1/p$, and $x = \log [-\log(1-F(\chi))]$. Given estimates of $F(\chi_i)$, say np_i, where i is the index of the i^{th} ordered Weibull random variable, the ordinary least squares analysis can be applied to estimate α and β. Using these estimates the least squares estimators are derived for the Weibull parameters $(\hat{\sigma}_0, \hat{p}_0)$ as $\hat{\sigma}_0 = \exp(\alpha)$ and $\hat{p}_0 = 1/\beta$. Several methods have been proposed for estimating $F(\chi_i)$. Following Engeman and Keefe (1982), $np_i = i/(n+1)$, is evaluated where i is the index of the i^{th} ordered Weibull random variable.

3.6.2 Monte Carlo simulation studies

The parameters of the Weibull distributions examined in this study are given in Table 3.7. The parameter values were selected to represent the range that may be observed in the study of air quality data. The sample sizes n=50, 100, 200 and 365 were chosen to represent the range of sample sizes which may be obtained when recording 24-h average data over a full year. For each sample size 5000 Monte Carlo experiments were undertaken with p = 1, 2, 3 and 4. The results are presented in Table 3.8.

Table 3.7: Parameters of the Weibull distributions used in the Monte Carlo experiments

b	c	$x_{0.01}$	$x_{0.99}$	Coefficient of variation
1.00	1.00	0.010	4.605	1.0000
1.00	2.00	0.100	2.146	0.5227
1.00	3.00	0.216	1.664	0.3690
1.00	4.00	0.317	1.465	0.2838

The results of Table 3.8 demonstrate in nearly all cases that the maximum likelihood estimators provide the best approximations to the upper percentiles of the Weibull distribution. All methods of parameter estimation are biased as indicated by the variation in the relative root mean square error with the percentile of the distribution.

The ordinary least squares estimators and the estimators derived by Menon (1963) would be preferred over the method of moments for deriving estimates of the upper percentiles of the Weibull distribution. The results of Table 3.8 indicate that $[\hat{x}_{0.99}]/x_{0.99}$, for the method of moments, remains significantly above all other methods for all sample sizes with the exception being where $\sigma = 1$, that is where the distribution is exponential. Given that the estimators derived by Menon (1963) may be rapidly computed, this method might be preferred over the ordinary least squares method and over the method of maximum likelihood for sample sizes larger than n = 365.

TABLE 3.8: Results of estimating the parameters of the Weibull distribution by four methods. The numbers are the result of 5000 Monte Carlo experiments.

	$[\hat{x}_{0.01}]/x_{0.01}$				$[\hat{x}_{0.50}]/x_{0.50}$				$[\hat{x}_{0.99}]/x_{0.99}$			
	Shape parameter				Shape parameter				Shape parameter			
Estimator	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000	1.000	2.000	3.000	4.000
Sample Size n = 50												
$(\hat{\sigma}, \hat{p})$	0.757	0.310	0.194	0.148	0.165	0.083	0.054	0.0416	0.187	0.094	0.063	0.047
$(\hat{\sigma}_m, \hat{p}_m)$	0.918	0.369	0.271	0.245	0.185	0.164	0.158	0.1600	0.210	0.149	0.145	0.147
$(\hat{\sigma}_e, \hat{p}_e)$	0.849	0.350	0.224	0.175	0.166	0.083	0.054	0.0417	0.285	0.133	0.087	0.066
$(\hat{\sigma}_o, \hat{p}_o)$	0.679	0.330	0.226	0.181	0.164	0.083	0.054	0.0419	0.348	0.153	0.099	0.075
Sample Size n = 100												
$(\hat{\sigma}, \hat{p})$	0.464	0.212	0.136	0.103	0.118	0.0585	0.0390	0.0294	0.133	0.066	0.044	0.033
$(\hat{\sigma}_m, \hat{p}_m)$	0.570	0.251	0.189	0.167	0.129	0.1130	0.1106	0.1105	0.151	0.104	0.102	0.102
$(\hat{\sigma}_e, \hat{p}_e)$	0.558	0.252	0.165	0.125	0.119	0.0588	0.0391	0.0295	0.195	0.093	0.062	0.046
$(\hat{\sigma}_o, \hat{p}_o)$	0.482	0.243	0.166	0.128	0.118	0.0587	0.0392	0.0296	0.224	0.103	0.068	0.051
Sample Size n = 200												
$(\hat{\sigma}, \hat{p})$	0.315	0.148	0.096	0.072	0.0836	0.0415	0.0280	0.0207	0.094	0.047	0.031	0.024
$(\hat{\sigma}_m, \hat{p}_m)$	0.388	0.175	0.133	0.115	0.0910	0.0794	0.0788	0.0774	0.105	0.073	0.072	0.072
$(\hat{\sigma}_e, \hat{p}_e)$	0.381	0.179	0.115	0.088	0.0839	0.0416	0.0281	0.0208	0.135	0.066	0.043	0.033
$(\hat{\sigma}_o, \hat{p}_o)$	0.347	0.175	0.116	0.090	0.0836	0.0415	0.0281	0.0208	0.149	0.071	0.046	0.036
Sample Size n = 365												
$(\hat{\sigma}, \hat{p})$	0.224	0.108	0.070	0.053	0.0609	0.0307	0.0203	0.0154	0.069	0.035	0.023	0.017
$(\hat{\sigma}_m, \hat{p}_m)$	0.279	0.127	0.096	0.085	0.0663	0.0584	0.0566	0.0571	0.078	0.054	0.052	0.053
$(\hat{\sigma}_e, \hat{p}_e)$	0.273	0.132	0.087	0.065	0.0611	0.0308	0.0204	0.0154	0.098	0.049	0.032	0.025
$(\hat{\sigma}_o, \hat{p}_o)$	0.257	0.131	0.088	0.067	0.0610	0.0308	0.0204	0.0154	0.105	0.052	0.034	0.026

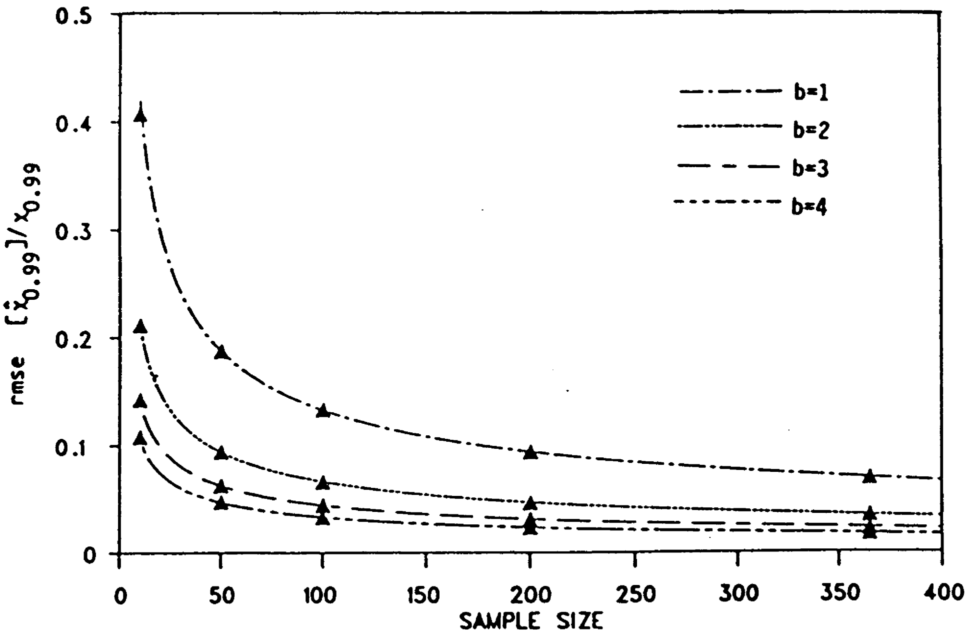

Figure 3.4: The fit accQrding to the empirical model (equation 3.46) to the rmse $[x_{0.99}]/x_{0.99}$ data of Table 3.8 for the maximum likelihood estimators of the Weibull distribution parameters.

3.6.3 The relative root mean square errors associated with the estimation of the percentiles of the Weibull distribution

The results of Table 3.8 give the minimum relative root mean square errors that may be expected when fitting the Weibull distribution to air quality data where the parameters of the distribution are unknown. For the 99-percentile an empirical model has been developed to provide an estimate of rmse $[\hat{x}_{0.99}]/x_{0.99}$, given by F(p,n). This empirical model is stated as

$$F(p,n) = \frac{\kappa}{p\surd n} \qquad (3.46)$$

where p is the Weibull shape parameter, n is the sample size, κ = 1.326 is a constant derived as the mean value of $\kappa = F(p,n)p\surd n$ using the data available in Table 3.8 for the maximum likelihood estimators. The fit according to equation (3.46) to the rmse $[\hat{x}_{0.99}]/x_{0.99}$ data for the maximum likelihood estimators is presented as Figure 3.4. It should be noted that a similar empirical model to that of equation (3.46) cannot be constructed for the estimators derived using the method of Menon (1963).

3.6.4 Conclusions for the Weibull distribution

The maximum likelihood estimators $\hat{\sigma}$ and $\hat{p}$ provide the best estimates of upper percentiles of the Weibull distribution. The estimators derived by Menon (1963), $\hat{\sigma}_e$ and $\hat{p}_e$, should provide the best alternative estimators to the maximum likelihood estimators especially at larger sample sizes ($n > 365$).

3.7 Autocorrelation and parameter estimation

In the preceding sections it has been assumed that the air quality observations are independently distributed random variables. However it is known that air quality observations do exhibit significant positive autocorrelations particularly for pollutants recorded over short time periods and pollutants exhibiting strong seasonal fluctuations, such as observations of ozone (Horowitz and Barakat, 1979; Hirtzel and Quon, 1981; Chock, 1984) where the autocorrelation, ρ, at lag 1 may be as high as 0.7.

The effect of autocorrelation upon the distribution of air quality observations does not change the expected value at a particular percentile but does increase the variance about that value (Chock, 1984). Hence based upon the assumption of no autocorrelation, the expected value at a particular percentile and the associated 95% confidence bound may indicate that air quality criteria will be violated only occasionally. However, the true variance may be consistent with a more frequent exceedance of the criteria.

In order to investigate the effect of autocorrelation, lognormal random deviates were generated with known autocorrelation (ρ= 0.25, 0.5, 0.75). The parameter values chosen for the lognormal distributions are listed in Table 3.1. Only the method of maximum likelihood has been considered in this study. Table 3.9 reports the rmse results based upon 1000 Monte Carlo experiments. Unfortunately a similar number of experiments (5000) to that of the original study of the lognormal distribution (the results of which are reported in Table 3.2) were not

Table 3.9: The rmse errors for the 99-percentile for autocorrelated lognormally distributed data using maximum likelihood parameter estimates.

lognormal parameter		Autocorrelation	
σ	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
		Sample size n=50	
0.1	0.0298	0.0339	0.0471
0.3	0.0933	0.1085	0.1445
0.5	0.1564	0.1727	0.2455
0.7	0.2111	0.2418	0.3340
		Sample size n=100	
0.1	0.0213	0.0249	0.0330
0.3	0.0637	0.0706	0.1056
0.5	0.1106	0.1261	0.1684
0.7	0.1529	0.1696	0.2409
		Sample size n=200	
0.1	0.0152	0.0175	0.0237
0.3	0.0437	0.0503	0.0699
0.5	0.0754	0.0877	0.1228
0.7	0.1068	0.1230	0.1646
		Sample size n=365	
0.1	0.0110	0.0128	0.0173
0.3	0.0338	0.0379	0.0518
0.5	0.0570	0.0663	0.0918
0.7	0.0773	0.0884	0.1275

possible due to the computational demands of generating autocorrelated sequences of data. Nevertheless the results of Table 3.9 remain sufficiently accurate for the purpose of demonstrating the effect of autocorrelation upon the variance of estimates of the upper percentiles. The results of Table 3.9 indicate that at $\rho = 0.25$ little difference between the rmse values of Table 3.9 and those of Table 3.2 can be observed. Even at $\rho = 0.5$ the rmse values have incressed only by about 25%. On the other hand the results where $\rho = 0.75$ indicate that for $\rho > 0.5$ underestimation of the variance in the estimates of the upper percentiles increases rapidly.

It was also found that the use of an empirical model such as that given by equation (3.14) could accurately describe the rmse values of Table 3.9 at each value of ρ. In each case a new value for the constant κ in equation (3.14) was required. These values were 2.17, 2.45 and 3.41 for autocorrelations of 0.25, 0.5 and 0.75 respectively. Estimates of the constant κ could be obtained at a range of autocorrelations allowing equation (3.14) to be applied to the full range of autocorrelation likely to be encountered in the study of air quality data.

For each of the data sets examined in the following chapters the autocorrelation coefficients were evaluated. Only in Chapter 7 for the 1-h and 0.5-h average recordings of sulphur dioxide, and in Chapter 6 for the 1-h average carbon monoxide data were the autocorrelations significant. Here the autocorrelations fell in the range from 0.25 to 0.75 indicating that the variance associated with model estimates may be larger than expected. The effect of this autocorrelation will be examined further in Chapters 6 and 7.

3.8 Conclusions

For all models considered here the method of maximum likelihood provides estimates of the upper percentiles with the lowest variance even at reasonably low sample size. However this is achieved at the expense of increased complexity of the estimation procedures and the need to obtain estimates using computationally demanding numerical methods. Alternative methods such as the method of moments may provide useful initial estimates of the pollutant concentration both for the purposes of air quality management and as initial guesses for the numerical procedures which evaluate maximum likelihood estimates. Where the increased uncertainty can be tolerated the methods providing direct analytical solutions may be preferred. Finally, it has been demonstrated that simple analytical formulae can easily be developed to yield reasonable approximations to the uncertainty associated with the estimation of the upper percentiles. These approximate confidence intervals represent the minimum uncertainty to be accounted for in any decision based upon estimates of the upper percentiles of the pollutant distribution.

CHAPTER 4
STATISTICAL IDENTIFICATION OF DISTRIBUTIONAL MODELS FOR AIR QUALITY CONCENTRATIONS

4.1 Introduction

In this chapter useful methods for the identification of statistical models are developed. An evaluation is presented of the relevance of standard goodness-of-fit tests to air pollution data using such criteria as power and robustness. Tests are applied to extensive pollutant data sets from Melbourne, Australia. Distribution type clearly depends on a number of factors: pollutant, time average, time period, emission source type (elevated, point, line, area), emission variation, meteorology and topography. If the effect of these factors on distribution type can be clarified then we are in a stronger position to infer distribution type when preparing impact assessments from new emission sources. Analysis of the Melbourne data set is a first step towards illuminating the contribution of these factors to distribution type.

4.2 Previous work

Many researchers have selected a priori a distribution to describe air pollution data. In a critical review of statistical distributions of air quality data, Georgopoulos and Seinfeld (1982) found for urban air quality data that this distribution is most likely to be the two-parameter lognormal. The decision to apply the lognormal distribution a priori can be attributed to the work of Larsen (1969, 1971, 1973, 1974) who, using graphical techniques, concluded that the upper percentiles of the distribution of air quality data are lognormally distributed for all pollutants, at all averaging times and for all urban sites considered. Clearly the need for a more objective assessment of goodness-of-fit, other than by a purely graphical approach, is required if the best distribution describing air quality data is to be determined. Graphical procedures based upon least squares analysis may however provide a useful measure in some calculations (Ott et al., 1979).

The earliest comprehensive study to report a comparison of the fit of several distributions to air quality data was carried out by

Bencala and Seinfeld (1976). In their study of 1-h average carbon monoxide data using the sum of squares error as the goodness-of-fit criteria, they found among the distributions considered, that a three-parameter lognormal distribution best described the pollutant distribution. In addition comparison of the results obtained by Bencala and Seinfeld (1976) when fitting the two-parameter gamma and lognormal distributions reveals that the gamma distribution yields a lower sum of squares error at half the sites studied. Other researchers have variously reported distributions such as the gamma (Pollack, 1975; Trijonis, 1978; Berger et al., 1982), exponential (Berger et al., 1982; Simpson et al., 1984) and Weibull distributions (Pollack, 1975) as applicable to the study of air quality data.

The problem then, given the range of alternatives available, and as no one distribution is conclusively the 'best', is to determine which distributional model is applicable to the data under study. Such a goodness-of-fit criterion should preferably be objective, as is also noted by Mage and Ott (1984), and should be applicable to a single data set. This last condition avoids recourse to the accumulation of a sum of squares error measured for many data sets from which the likely distributional form may be inferred. While such information is clearly important in the assessment of goodness-of-fit, and should prove useful in assessing the error associated with such procedures, other statistical tests are available. Some tests may be termed 'traditional' tests of goodness-of-fit of the hypothesized distribution to the observations. Such statistical tests satisfy the requirements that the measurement of goodness-of-fit should be objective and should provide a means by which the results of one analysis for distributional form may be directly compared with another. These statistical tests also indicate when the hypothesis of a particular distributional form for air quality data may be rejected. Thus models which are unsatisfactory may be readily identified and new models sought.

Finally, traditional goodness-of-fit tests take into account the variability associated with sampling from a particular distributional form. Confidence levels may be readily determined. Thus, for example, the natural fluctuations in sampling from a lognormal distribution can be accounted for, while variation outside the expected range can be identified.

4.3 Model identification - testing goodness-of-fit

When examining air quality data with the view to identifying a suitable distributional model, two avenues of investigation are available. In the first case methods can be developed based upon the observed data, to test the hypothesis, say, that the sampled population of air quality data is normal, lognormal, exponential, gamma, Weibull or any other distribution. These are the goodness-of-fit methods. There is also the measure of robustness which refers to how far the distribution of observations may differ from the hypothesised distribution before substantial errors arise in the percentile estimates derived theoretically from a particular distributional form.

In this chapter the methods by which goodness-of-fit may be evaluated are considered. Robustness is not examined here. However it is likely that the robustness required will vary with the intended application and that some measure of goodness-of-fit will be needed to assess whether the distributional model is sufficiently robust for that application.

Using methods of statistical inference, the hypothesis, which is designated as the null hypothesis (H_0), that air pollution data are lognormally distributed will be tested against the alternative hypothesis (H_1), that the data follow some other distributional form. When testing the null hypothesis the test can be classified as either simple or composite. The hypothesis is simple when testing for a lognormal model with the parameters of the distribution known and thus only one probability function is defined on the sample space. Where the value of the parameters of the lognormal distribution must be estimated from the sample, the hypothesis is termed composite.

In this study the parameters of the lognormal, exponential, Weibull and gamma distributions are estimated from the sample, and thus the hypotheses tested will be composite. When testing a composite hypothesis the range of tests available is restricted and the ability of these tests to distinguish between data drawn from, for example, the lognormal distribution, from that drawn from some other distributional form, is reduced. It is advisable to examine the performance of the goodness-of-fit tests using Monte Carlo techniques under a variety of conditions in order to ascertain whether the test is suitable for the desired application and to determine which test is likely to be the most powerful. It should be noted that the power of a statistical test is the probability of rejecting a false null hypothesis (Conover, 1980). Hence a statistical test with low power when testing for lognormality would accept data drawn from, for example, a gamma distribution, with a high probability.

One further restriction imposed upon the range of goodness-of-fit tests available is the range of sample sizes to which the tests may be applied. Particular attention has been given to testing small sample sizes, typically $n < 50$, with several tests for the normal distribution available for this range of sample sizes. With regard to sample size, it should also be noted that with an increasing sample size of air pollution data the probability that a theoretical distribution will fit this data exactly will decline (Conover, 1980). This problem arises from the conflict between the limited number of theoretical distributions available and the infinite number of distributions observable in nature (Conover, 1980). A goodness-of-fit test statistic will still provide a useful relative measure of fit when comparing alternative distributional models. Using the test statistic a model may be selected or new models may be sought. In any case, where new models are considered some measure of goodness-of-fit will be required to determine whether an improved fit has resulted.

4.4 Goodness-of-fit tests

Let $\chi_1, \chi_2, \ldots, \chi_n$ be independent observations of a random variable with distribution function $F(\chi)$ which is unknown. The goodness-of-fit problem is then to test the hypothesis

$$H_o : F(\chi) = F_o(\chi) \tag{4.1}$$

where $F_o(\chi)$ is a particular distribution function which in this case may be continuous. This study examines the chi-square test, Kolmogorov-Smirnov test, a test statistic of the Kolmogorov type developed by

Lilliefors (1967), a test statistic developed by Shapiro and Wilk (1965), a test statistic developed by Shapiro and Francia (1972) and a test statistic developed by D'Agostino (1971, 1972).

Given a set of air pollution observations x_i (i = 1 ... n) the chi-square statistic X^2 may be stated as follows (Kendall and Stuart, 1973)

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - np_i]^2}{np_i} \qquad (4.2)$$

where k is the number of equiprobable cells in which the total number of observations n, may fall, np_i is the expected number of observations in each cell, and n_i the number of observations falling in that cell. When using X^2 as a test statistic, H_0 is to be rejected when X^2 is large. As X^2 has a chi-squared distribution asymptotically with (k-1) degrees of freedom we may determine the probability of not exceeding X^2. Where the probability α falls below 0.05, we reject the hypothesis H_0 at the 95% confidence level. In order to test the composite hypothesis the number of degrees of freedom should be reduced by the number of parameters estimated from the sample. Thus, using the chi-square test, the composite hypothesis may be readily tested.

Unfortunately when applying the chi-square test problems can be encountered with the selection of the number of equiprobable cells, k. Goodness-of-fit is tested for only these k classes rather than the n observations (Gibbons, 1971; Kendall and Stuart, 1973; Conover, 1980). As an example of the application of the chi-square test to air quality data Tong and De Pietro (1977) used the chi-square test to examine 24-h average sulphate particulates and found at the 99% confidence level that the hypothesis of lognormality could be rejected at only one of the twelve sites at which data were collected. Other distributional models were not considered. That lognormality was accepted at most sites does not imply that better models cannot be found.

There exists a more general test of fit proposed by Kolmogorov. It is based on the cumulative distribution of the sample, or sample distribution function, which we define as

$$S_n(\chi) \begin{cases} = o & \chi < \chi_{(1)} \\ = \frac{r}{n} & \chi_{(r)} < \chi < \chi_{(r+1)} \\ = 1 & \chi_{(n)} < \chi \end{cases} \tag{4.3}$$

The χ_r are the order statistics, that is, the observations arranged so that

$$\chi_1 < \chi_2 < \dots < \chi_n \tag{4.4}$$

Thus $S_n(\chi)$ is the proportion of the observations not exceeding χ. If $F_o(\chi)$ is the true distribution function, fully specified, from which the observations are drawn, we have, for each value of χ, from the Law of Large Numbers (Kendall and Stuart, 1973)

$$\lim_{n\to\infty} P \{ S_n(\chi) = F_o(\chi)\} = 1 \tag{4.5}$$

The test statistic proposed by Kolmogorov is based on deviations of the sample distribution function $S_n(\chi)$ from the completely specified continuous hypothetical distribution function $F_o(\chi)$. The measure of deviation is the maximum absolute difference between $S_n(\chi)$ and $F_o(\chi)$ defined as

$$D_n = \sup_{\chi} |S_n(\chi) - F_o(\chi)| \tag{4.6}$$

The Kolmogorov statistic is regarded as one of the most important of the general tests of fit as an alternative to χ^2 (Kendall and Stuart, 1973). As the distribution of D_n is known, again we may determine the probability, α, of not exceeding D_n. At the 95% confidence level, where the probability α falls below 0.05, we reject the hypothesis H_o. Kalpasanov and Kurchatova (1976) have applied the

Kolmogorov statistic to 24-h average lead particulates, phenol, hydrogen cyanide, nitric oxide, nitrogen dioxide, sulphur dioxide, hydrogen sulphide, formaldehyde and oxidants and found at a 95% confidence level that the data were not lognormally distributed. Other distributional models were not considered in this study.

Unfortunately the standard tables used in conjunction with the Kolmogorov test are only valid when testing whether a set of observations were drawn from a completely specified continuous distribution. If one or more parameters must be estimated from the sample then the tables are no longer valid. The effect of using these tables when parameters must be estimated from the sample is to reduce the probability of rejecting the null hypothesis. Accordingly Lilliefors (1967) developed a statistical test for normality with the test statistic defined as

$$D = \sup \left|S_n(x) - F^*(x)\right| \qquad (4.7)$$

where $F^*(x)$ is the cumulative distribution function with the parameters estimated using sample mean and variance. Lilliefors (1967) presents a tabulation of the critical values of D.

The test statistic developed by Shapiro and Wilk (1965) is confined to the normal and by transformation the lognormal distributions. It is calculated as

$$W = \{\sum_{i=1}^{h} a_{in} (x_{(n-i+1)} - x_{(i)})\}^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4.8)$$

where $h = \frac{1}{2}n$ or $\frac{1}{2}(n-1)$ according to whether n be even or odd. Shapiro and Wilk (1965) give a table of coefficients a_{in} for $n = 2(1)50$, $i=1(1)h$. This test is thus restricted to a useful range up to a sample size of n=50. As the test of Shapiro and Wilk (1965) does not extend beyond n=50 Shapiro and Francia (1972) have suggested a very similar criterion, W', with which to replace W if n > 50. This test statistic is given as

$$W' = \{\sum_{i=1}^{h} b_{in} (x_{(n-i+1)} - x_i)\}^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4.9)$$

where the coefficients b_{in} are derived from the expected normal order statistics. Shapiro and Franica (1972) give the lower and upper significance levels, 1, 5, 10, 15, 20% for W' for n = 35, 50, 51(2)99. In this study we shall not be examining the power of these tests as they are applicable to only small sample sizes. However, it is worth noting that comparative studies of the power of various tests for normality (Shapiro, Wilk and Chen, 1968; Pearson, D'Agostino and Bowman, 1977) have indicated that the W test is the most powerful test against a wide range of alternative tests including those considered in this paper. Thus at small sample sizes this test would be preferred if there is no practical problem with incorporating tables of a_{in} coefficients in a computer program.

The test statistic developed by D'Agostino (1971), which is denoted by Y, is given by

$$Y = n^{1/2} \left(\frac{D-0.282095}{0.0299860}\right) \qquad (4.10)$$

where

$$D = \sum_{i=1}^{h} C_{in}(x_{(n-i+1)} - x_i)/[n^{3/2}\{\sum_{i=1}^{n} (x_i - \bar{x})^2\}^{1/2}] \qquad (4.11)$$

and the coefficients C_{in} adopt the simple linear form

$$C_{in} = 1/2\,(n+1) - i \qquad (4.12)$$

D'Agostino (1971) has tabulated the upper and lower 0.5, 1.0, 2.5, 5.0 and 10.0 percentiles of the distribution of the test statistic Y for n=100(50)1000 and, D'Agostino (1972) has prepared similar tables for the small sample sizes n=2(2)50(10)100 at the same percentage points. This test is applied as a two-sided test and thus the upper and lower percentage points are required.

In order to assess the ability of the chi-square test, the Lilliefors test and the D'Agostino test to distinguish samples drawn from non-normal populations, a simulation study was performed. A wide range of non-normal populations have been examined. Table 4.1 lists the systems of population distributions and their equations. The range of distributions and populations sampled is comparable in range to that used by Shapiro, Wilk and Chen (1968) and Pearson, D'Agostino and Bowman (1977). In this thesis the power of these test statistics are examined at a sample size of n=365. This sample size was chosen to reflect a sample of 24-h averaged data recorded over one year. The estimates of power, based upon 100 Monte Carlo replications for the populations sampled, are listed in Table 4.2.

The results of Table 4.2 indicate that the test developed by Lilliefors (1967) is more powerful in all cases than the chi-square test. Similarly the D'Agostino statistic provides a more powerful test than that provided by the chi-square test. However comparison with the Lilliefors statistic does not clearly distinguish it as more powerful. The D'Agostino statistic does not appear to reject location contaminated normal populations. In practice such distributions might occur where a change in the mean level of emissions has occurred. On the basis of the range of non-normal populations considered, the Lilliefors test appears to provide the most powerful test from amongst those considered in this study for samples of size greater than n=100. The Lilliefors test also has the advantage that it may be readily applied to all sample sizes, as estimates of the significance levels are available for all $n > 3$, which covers the full range of sample sizes likely to be observed in air quality studies.

4.5 Selecting a distributional model from among exponential, gamma, lognormal and Weibull alternatives

In the preceding section various test statistics for evaluating the goodness-of-fit of the normal, and by transformation of the data, lognormal distributions to air quality observations were examined. It was found that a modification of the Kolmogorov test developed by Lilliefors (1967) provides a reasonably powerful goodness-of-fit test.

Table 4.1: Equations of the distributions sampled for the study of the power of the tests of goodness-of-fit

Distribution	Equation	Range
Gamma or χ^2 (p is an even integer)	$f(\chi) = \{\Gamma(p)\}^{-1}\chi^{p-1}e^{-\chi}$	$(0 < \chi < \infty)$
Lognormal	$f(\chi) = \{\chi\sigma(2\pi)^{\frac{1}{2}}\}^{-1}e^{-(\ln\chi)^2/2\sigma^2}$	$(0 < \chi < \infty)$
Weibull	$f(\chi) = \kappa\chi^{\kappa-1}e^{-\chi^{\kappa}}$	$(0 < \chi < \infty;\ \kappa > 0)$
Logistic	$f(\chi) = \beta e^{\beta\chi}/(1+e^{\beta\chi})^2$	$(-\infty < \chi < \infty)$
Scale contaminated[a]	$f(\chi) = \{(1-p)e^{-\frac{1}{2}\chi^2} + (p/\lambda)e^{-\frac{1}{2}\chi^2\lambda^{-2}}\}(2\pi)^{-\frac{1}{2}}$	$(-\infty < \chi < \infty)$
Location contaminated[b]	$f(\chi) = \{(1-p)e^{-\frac{1}{2}\chi^2} + pe^{-\frac{1}{2}(\chi-\mu)^2}\}(2\pi)^{-\frac{1}{2}}$	$(-\infty < \chi < \infty)$
Exponential	$f(\chi) = \cancel{1\ e^{(-\chi/b)}}\ (1/b)\exp(-x/b)$	$(0 < \chi < \infty)$
Beta	$f(\chi) = \{B(a,b)\}^{-1}\chi^{a-1}(1-\chi)^{b-1}$	$(0 < \chi < 1)$
Student's t	$f(\chi) = \Gamma\{\frac{1}{2}(\nu+1)\}\{\Gamma(\frac{1}{2}\nu)\}^{-1}(\nu\pi)^{-\frac{1}{2}}(1+\chi^2/\nu)^{-\frac{1}{2}(\nu-1)}$	$(-\infty < \chi < \infty)$

[a] The scale contaminated distribution is composed of two superimposed normal curves having the same means but differing standard deviations.

[b] The location contaminated distribution is composed of two superimposed normal curves having the same standard deviations but different means.

Table 4.2: Estimates of power based on 100 replications each of size n=365. The number of samples rejected is given.

Population Sampled	Goodness-of-fit Test		
	Chi-square	Lilliefors	D'Agostino
Weibull $\kappa=0.5$	100	100	100
$\kappa=2.0$	65	97	24
$\kappa=3.0$	2	10	24
Beta a=1, b=1	100	100	100
a=2, b=2	64	80	100
a=3, b=2	60	90	78
a=2, b=1	100	100	7
Chi-square $\nu=10$	78	99	83
$\nu=4$	100	100	100
$\nu=1$	100	100	100
Student's t $\nu=6$	17	68	95
$\nu=4$	44	91	100
$\nu=2$	100	100	100
$\nu=1$	100	100	100
Logistic $\beta=1$	12	44	75
Exponential b=1	100	100	100
Gamma p=1.5	100	100	100
p=0.8	100	100	100
SC;p=0.05, $\lambda=3$	23	67	97
SC;p=0.10, $\lambda=3$	47	90	100
SC;p=0.20, $\lambda=3$	89	100	100
SC;p=0.05, $\lambda=5$	91	99	100
SC;p=0.10, $\lambda=5$	100	100	100
SC;p=0.20, $\lambda=5$	100	100	100
SC;p=0.05, $\lambda=7$	100	100	100
SC;p=0.10, $\lambda=7$	100	100	100
SC;p=0.20, $\lambda=7$	100	100	100
LC;p=0.05 $\mu=3$	51	87	96
LC;p=0.10, $\mu=3$	85	100	99
LC;p=0.20, $\mu=3$	94	100	72
LC;p=0.30, $\mu=3$	96	100	7
LC;p=0.05, $\mu=5$	100	100	100
LC;p=0.20, $\mu=5$	100	100	100
LC;p=0.05, $\mu=7$	100	100	100
LC;p=0.10, $\mu=7$	100	100	100
Lognormal $\mu=3,\sigma=1.0$	100	100	100
$\mu=3,\sigma=0.5$	100	100	100
$\mu=3,\sigma=0.3$	82	100	89
$\mu=3,\sigma=0.2$	37	85	48
$\mu=3,\sigma=0.1$	10	26	7

LC = location contaminated; SC=scale contaminated

In this section the problem of identifying an appropriate distributional form for a random variable given a random sample of observations of it is considered. This study is motivated by the need to identify the appropriate distributional form with which to model air quality data. For this application it is necessary to assume that air pollutant concentrations are independent, identically distributed random variables. That this is a reasonable assumption is confirmed by Georgopoulos and Seinfeld (1982) who in their review of the statistical distributions of air pollutant concentrations note that the application of theoretical results derived for independent, identically distributed random variables produces satisfactory agreement with observations.

From the study of air quality data collected over a fixed averaging time, common distributional forms considered appropriate are the two-parameter lognormal, gamma and Weibull distributions (Bencala and Seinfeld, 1976). The exponential distribution has also received interest as a model of pollutant concentration observations about isolated point sources (Simpson et al., 1984). In general then it is reasonable to assume that the distributions describing air quality data are unimodal and skewed to the right.

An important goal in the evaluation of air quality is estimation of the upper percentiles of pollutant distributions. These percentiles are of interest as many air quality standards are written in terms of the frequency with which a particular level may be exceeded. For example, many standards must not be exceeded more than once per year and thus accurate estimation of the second highest pollutant concentration can be of critical importance when comparing air quality observations with standards.

Additional to the interest in the upper percentiles is the need to accurately estimate the entire range of the distributions of pollutant concentrations. Such information may be applicable for the analysis of damages sustained. The full range of the distribution is of interest as significant damage, for example to materials, might also be produced as a result of low concentrations occurring with high frequency. The lower concentrations may also be of importance where a synergistic combination of pollutants occurs.

What is required then, is a procedure by which a distributional model may be selected which best represents the entire distribution of pollutant concentrations from amongst a reasonable range of alternatives. The results obtained here of course are general and extend to situations other than where the random variable is pollutant concentration over a given averaging time.

Bain and Engelhardt (1980) examined a procedure, based on the evaluation of the logarithm of the likelihood function, for selecting between the Weibull and gamma distributions, one of which being the true distribution. The procedure involves selection of the distributional form which yields the largest value of the gamma and Weibull log likelihood functions where the parameters of the likelihood functions are estimated using the method of maximum likelihood. Simulation studies indicated that the maximized log likelihood function provides an excellent basis for selecting between the Weibull and gamma distributions.

More recently Kappenman (1982) extended the work of Bain and Engelhardt (1980) to include the lognormal distribution. In this study the results of selecting the model which produced the largest logarithm of the maximimum likelihood function showed that even where three distributions are considered a high probability existed of selecting the correct distribution.

However when considering air quality data it is unlikely that only three models of the distribution will adequately describe all sets of observations. In particular, and as noted earlier, the exponential model should be included in such an analysis. Including such a model is of practical importance as it is desirable to model air quality data especially, with the smallest number of parameters. Conversely it is necessary to determine when a 2-parameter model may be more appropriate than the 1-parameter exponential model. Such a result is of importance in the study of isolated point sources where resultant ambient concentrations can be exponential.

In addition to achieving model parsimony where appropriate, it may also be desirable to reject the hypothesis of the observations belonging to one of the distributional models selected where in fact none of the models are adequate. Bain and Engelhardt (1980) do not choose a significance level with which to apply their tests as the assumption is made that one of the distributions is the correct one. Effectively then the probability of making a type I error as well as rejecting the remaining distributional models is zero.

In order to provide a test of the appropriateness of the distributional models, test statistics of the Kolmogorov type were employed, where account is taken for the estimation of the parameters from the sample using the method of maximum likelihood (Lilliefors, 1967). This test was chosen because of its high power and ease of application but clearly other tests of goodness-of-fit where a confidence level may be stated could be substituted. The commonly used chi-squared test for example, does have the disadvantages that the number and character of class intervals used is arbitrary and for the smaller samples the number of cells in the chi-square test must also be small.

As a comparison with the procedure which selects the distribution with the largest value of the likelihood function, the ratio of the Kolmogorov test statistic to a given confidence level is calculated. Using such a ratio, the distributional model with the lowest value of this ratio is selected. The 95% confidence level was chosen to provide a reasonably low probability of making a type I error while retaining the ability to reject an incorrect hypothesis and thus allowing new models to be developed where necessary. Depending upon the objectives of the model identification, other confidence levels could also be useful. A range of confidence levels could be investigated by examining the errors associated with the application of models accepted at various confidence levels, however this problem is not examined here.

To assess the performance of these two selection procedures a simulation study was performed. For each exercise the choice was among Weibull, gamma, exponential and lognormal forms with the true distribution being one of these. To provide a direct comparison with the likelihood ratio selection criterion, the distributional model with the minimum value of the Kolmogorov ratio is selected without rejecting distributional models which are not accepted at the 95% confidence level.

4.5.1 The selection criteria

The probability density functions for the two parameter lognormal, gamma, Weibull and exponential distributions are, respectively, of the form

$$f_1(\chi) = (1/\sqrt{2\pi}c) \exp \{-[\ln (\chi/b)]^2/2c^2\} \tag{4.13}$$

$$f_2(\chi) = [1/b^c\Gamma(c)]\ \chi^{c-1} \exp (-\chi/b) \tag{4.14}$$

$$f_3(\chi) = (c/b)\ (\chi/b)^{c-1} \exp [-\ (\chi/b)^c] \tag{4.15}$$

$$f_4(\chi) = (1/b) \exp (-\ \chi/b) \tag{4.16}$$

where in all cases c represents the shape parameter and b the scale parameter. Now if $\chi_1, \chi_2, \ldots, \chi_n$ represents a random sample of n observations of say, air quality data, then the logarithms of the maximum likelihood functions for the lognormal, gamma, Weibull and exponential distributions are respectively

$$\ln L_1 = n \ln \left(\frac{1}{c\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \ln \chi_i - \frac{1}{2c^2} \sum_{i=1}^{n} (\ln \chi_i - \ln b)^2 \tag{4.17}$$

$$\ln L_2 = -nc \ln b - n \ln \Gamma(c) + (c-1) \sum_{i=1}^{n} \ln (\chi_i) - \sum_{i=1}^{n} \chi_i/b \tag{4.18}$$

$$\ln L_3 = n (\ln c - \ln b) + (c-1) \sum_{i=1}^{n} \ln (\chi_i) - \sum_{i=1}^{n} \chi_i^c \tag{4.19}$$

$$\ln L_4 = -n \ln b - \frac{1}{b} \sum_{i=1}^{n} \chi_i \tag{4.20}$$

Here Γ is the gamma function. The parameters b and c for each likelihood function were calculated using the method of maximum likelihood (see Chapter 3).

The distributions of test statistics were obtained by Monte Carlo simulation for all distributions in this thesis excepting the exponential distribution for which tables of the exact distribution are available (Durbin, 1975). The test statistics are of the Kolmogorov type and may be evaluated as

$$T_n = \sup_{\chi} \left| F(\chi) - S(\chi) \right| \qquad (4.21)$$

where $S(\chi)$ is the empirical distribution function and $F(\chi)$ is the distribution function for the lognormal, gamma and Weibull models. Using 10,000 Monte Carlo experiments and estimating the parameters of $F(\chi)$ using the method of maximum likelihood the 95% confidence levels were obtained at n = 4(1)30. Using these data, approximations to the 95% confidence levels at n=50, 200 and 365 were derived (for example Lilliefors, 1967). For the gamma distribution, however, a single table of the test statistic is not obtainable. The distribution of this test statistic changes with the value of the shape parameter since it is not possible to transform the gamma distribution to a form independent of the value of the shape parameter (Johnson and Kotz, 1970). Fortunately the critical value undergoes only a small change over the range of values of the shape parameter (c = 1-8) of interest in the study of air quality. Larger variations do occur where the shape parameter is less than unity. The procedure adopted then was to use the maximum value of the test statistic over this range thus slightly increasing the risk of a type II error. For the gamma distribution this allows comparison of the ratio of the Kolmogorov statistic to a confidence level with that obtained for the other distributional models.

In order to select the distributional model from among the Weibull, gamma, exponential and lognormal models the distribution with the minimum value of the ratio of the test statistic T_n to the respective 95% confidence level D_{95} or in the case of the gamma distribution its approximate 95% confidence level, is selected. The procedure based on the log likelihood function selects the distributional model yielding the maximum value of the log likelihood function.

4.5.2 Simulation procedure and results

To assess the performances of these procedures estimates of the probabilities of the correct selection have been obtained by simulation over a range of possible cases which may arise in the study of air quality. The probability of correct selection can be shown to be independent of the value of the scale parameter b, without regard for the underlying distribution (Kappenman, 1982), however it does vary with the value of the shape parameter. Accordingly 1000 simulations of sample size n = 10 and 25, 500 simulations with a sample size n = 50 and 100, 250 simulations with sample size n = 200 and 365, were undertaken

(i) from lognormal distributions with shape parameter c = 0.5, 1.0, 2.0, 3.0, 4.0,

(ii) from gamma distributions with shape parameter c = 1.0, 2.0, 3.0, 4.0,

(iii) from Weibull distributions with shape parameter c = 1.0, 2.0, 3.0, 4.0 and,

(iv) from an exponential distribution with scale parameter b = 1.

For each of the above sample size-shape parameter combinations, the probabilities of selecting the exponential, Weibull, gamma and lognormal models using both criteria were evaluated. As a first step, and in order to verify the operation of the simulation program, the results of Kappenman (1982) were examined for the case where the three models Weibull, gamma and lognormal were selected on the basis of the maximum value of the log likelihood function. Agreement was found except in the case of the gamma distribution. In contrast to Kappenman (1982), very low values for the probability of correctly selecting the gamma model did not result. A possible explanation for obtaining unusually low probabilities is that the estimate of the scale parameter may not have been included in the evaluation of the log likelihood function. Where the estimate of the scale parameter is included in the evaluation of the gamma log likelihood function the probability of correct selection of the gamma model rises to similar levels obtained for the Weibull and lognormal models.

Results of the Monte Carlo experimentation are presented for the case where the exponential distribution is the underlying distribution in

Table 4.3; for the gamma distribution with c = 2.0 in Table 4.4; for the Weibull distribution with c = 2.0 in Table 4.5; and for the lognormal distribution with c = 0.5, 1.0 in Tables 4.6 and 4.7. Figure 4.1 presents the probability of selecting a particular model where the underlying distribution is exponential when applying the criterion based on the minimum value of the ratio of the Kolmogorov type test statistic to the respective confidence interval. The results of the log likelihood procedure are not displayed, as from Table 4.3 the probability of selecting the exponential distribution using this criterion is zero over all the sample sizes considered. Figure 4.2 presents for the gamma distribution the probability of selecting the correct distribution using both the log likelihood and Kolmogorov selection criteria. Figures 4.3 and 4.4 present similar plots for the Weibull and lognormal distributions.

Examining Table 4.3 it should be noted that for all sample sizes that the log likelihood criterion fails to select the exponential distribution as the appropriate model. This is not the case however for the Kolmogorov based selection criterion. Figure 4.1 illustrates the behaviour of this selection criterion. This test selects the exponential model with high probability over the majority of the range of sample sizes. It should also be noted that this procedure selects the lognormal model, an important alternative model applicable to point source air pollution observations (Simpson et al., 1984), with very low probability for sample sizes above n = 50. Selection of the Weibull and gamma alternatives does not appear to decline with increasing sample size reflecting the fact that the exponential distribution is a special case of both the Weibull and gamma distributions.

Table 4.3: Percentage of distributions selected where the underlying distribution is exponential. The first column gives, for each model, the percentage accepted at the 95% confidence level using the Kolmogorov based test statistics. The second and third columns represent the frequency with which each model was selected as the best distributional model from amongst the listed alternatives.

Distribution	% Accepted	Maximum log likelihood	Minimum T_n/D_{95}
	Sample size n = 10		
lognormal	94.1	30.4	24.0
exponential	96.4	0.0	52.6
gamma	95.9	55.5	7.7
Weibull	97.0	14.1	15.7
	Sample size n = 25		
lognormal	78.4	16.5	12.9
exponential	95.6	0.0	58.9
gamma	92.7	64.1	11.1
Weibull	94.7	19.4	17.1
	Sample size n = 50		
lognormal	54.8	8.4	8.2
exponential	95.6	0.0	63.6
gamma	91.4	66.8	8.4
Weibull	94.0	24.8	19.8
	Sample size n = 100		
lognormal	25.2	3.6	4.0
exponential	94.4	0.0	66.8
gamma	90.6	66.6	11.4
Weibull	93.0	29.8	17.8
	Sample size n = 200		
lognormal	2.8	0.8	0.4
exponential	93.6	0.0	70.0
gamma	90.4	71.2	14.4
Weibull	92.4	28.0	15.2
	Sample size n = 365		
lognormal	0.4	0.0	0.0
exponential	96.8	0.0	65.6
gamma	94.0	48.0	16.0
Weibull	92.9	52.0	18.4

Table 4.4: Percentage of distributions selected where the underlying distribution is gamma with c = 2.0. The first column gives, for each model, the percentage accepted at the 95% confidence level using the Kolmogorov based test statistics. The second and third columns represent the frequency with which each model was selected as the best distributional model from amongst the listed alternatives.

	%	Maximum	Minimum
Distribution	Accepted	log likelihood	T_n/D_{95}
Sample size n = 10			
lognormal	94.6	35.4	35.7
exponential	83.1	0.0	22.8
gamma	94.6	61.0	15.7
Weibull	96.4	3.6	25.8
Sample size n = 25			
lognormal	86.3	25.0	28.6
exponential	56.7	0.0	6.2
gamma	95.0	74.2	32.3
Weibull	92.9	0.8	32.9
Sample size n = 50			
lognormal	76.8	18.8	22.8
exponential	15.0	0.0	1.4
gamma	94.4	91.2	43.0
Weibull	89.2	0.0	32.8
Sample size n = 100			
lognormal	54.4	7.6	16.0
exponential	0.6	0.0	0.0
gamma	94.8	92.4	52.0
Weibull	87.2	0.0	32.0
Sample size n = 200			
lognormal	26.0	2.0	6.4
exponential	0.0	0.0	0.0
gamma	93.2	98.0	68.4
Weibull	80.0	0.0	25.2
Sample size n = 365			
lognormal	5.2	0.0	1.2
exponential	0.0	0.0	0.0
gamma	94.4	100.0	81.6
Weibull	70.8	0.0	17.2

Table 4.5: Percentage of distributions selected where the underlying distribution is Weibull with c = 2.0. The first column gives, for each model, the percentage accepted at the 95% confidence level using the Kolmogorov based test statistics. The second and third columns represent the frequency with which each model was selected as the best distributional model from amongst the listed alternatives.

Distribution	% Accepted	Maximum log likelihood	Minimum T_n/D_{95}
		Sample size n = 10	
lognormal	92.7	22.5	33.8
exponential	58.5	0.0	8.4
gamma	94.3	32.4	12.7
Weibull	96.2	45.1	45.1
		Sample size n = 25	
lognormal	79.4	9.8	19.5
exponential	9.7	0.0	1.2
gamma	91.8	32.8	22.6
Weibull	94.8	57.4	56.7
		Sample size n = 50	
lognormal	56.8	3.6	8.2
exponential	0.2	0.0	0.2
gamma	85.0	30.2	25.4
Weibull	95.4	66.2	66.2
		Sample size n = 100	
lognormal	27.0	0.2	2.0
exponential	0.0	0.0	0.0
gamma	74.8	22.2	25.8
Weibull	93.8	77.6	72.2
		Sample size n = 200	
lognormal	3.6	0.0	0.0
exponential	0.0	0.0	0.0
gamma	49.2	8.8	13.6
Weibull	94.0	91.2	86.4
		Sample size n = 365	
lognormal	0.0	0.0	0.0
exponential	0.0	0.0	0.0
gamma	40.0	7.6	10.4
Weibull	93.6	92.4	89.6

Table 4.6: Percentage of the distributions selected where the underlying distribution is lognormal with c = 0.5. The first column gives, for each model, the percentage accepted at the 95% confidence level using the Kolmogorov based test statistics. The second and third columns represent the frequency with which each model was selected as the best distributional model from amongst the listed alternatives.

Distribution	% Accepted	Maximum log likelihood	Minimum T_n/D_{95}
	Sample size n = 10		
lognormal	96.4	58.8	68.1
exponential	31.6	0.0	14.0
gamma	93.7	41.2	10.4
Weibull	93.4	0.0	20.1
	Sample size n = 25		
lognormal	95.5	65.4	68.7
exponential	0.1	0.0	0.0
gamma	89.0	34.6	16.9
Weibull	80.1	0.0	14.4
	Sample size n = 50		
lognormal	94.4	73.2	72.6
exponential	0.0	0.0	0.0
gamma	81.8	26.8	20.0
Weibull	58.4	0.0	7.4
	Sample size n = 100		
lognormal	94.8	82.0	79.6
exponential	0.0	0.0	0.0
gamma	71.0	18.0	18.4
Weibull	31.6	0.0	2.0
	Sample size n = 200		
lognormal	94.2	92.8	85.2
exponential	0.0	0.0	0.0
gamma	48.8	7.2	14.4
Weibull	4.8	0.0	0.4
	Sample size n = 365		
lognormal	95.6	96.8	90.0
exponential	0.0	0.0	0.0
gamma	33.6	3.2	10.0
Weibull	0.0	0.0	0.0

Table 4.7: Percentage of distributions selected where the underlying distribution is lognormal with c = 1.0. The first column gives, for each model, the percentage accepted at the 95% confidence level using the Kolmogorov based test statistics. The second and third columns represent the frequency with which each model was selected as the best distributional model from amongst the listed alternatives.

Distribution	% Accepted	Maximum log likelihood	Minimum T_n/D_{95}
	Sample size n = 10		
lognormal	95.1	66.2	47.8
exponential	92.3	0.0	36.7
gamma	86.8	33.8	6.2
Weibull	91.9	0.0	9.3
	Sample size n = 25		
lognormal	96.3	77.1	62.4
exponential	85.7	0.0	26.2
gamma	72.4	22.9	5.7
Weibull	79.2	0.0	5.7
	Sample size n = 50		
lognormal	95.8	86.8	74.4
exponential	76.4	0.0	17.0
gamma	53.6	13.2	5.2
Weibull	61.4	0.0	3.4
	Sample size n = 100		
lognormal	94.2	95.6	85.8
exponential	55.4	0.0	10.2
gamma	24.8	4.4	2.8
Weibull	31.6	0.0	1.2
	Sample size n = 200		
lognormal	94.4	98.8	94.8
exponential	15.6	0.0	2.8
gamma	3.6	1.2	2.0
Weibull	7.6	0.0	0.4
	Sample size n = 365		
lognormal	94.0	100.0	98.8
exponential	0.0	0.0	0.8
gamma	0.0	0.0	0.0
Weibull	0.0	0.0	0.4

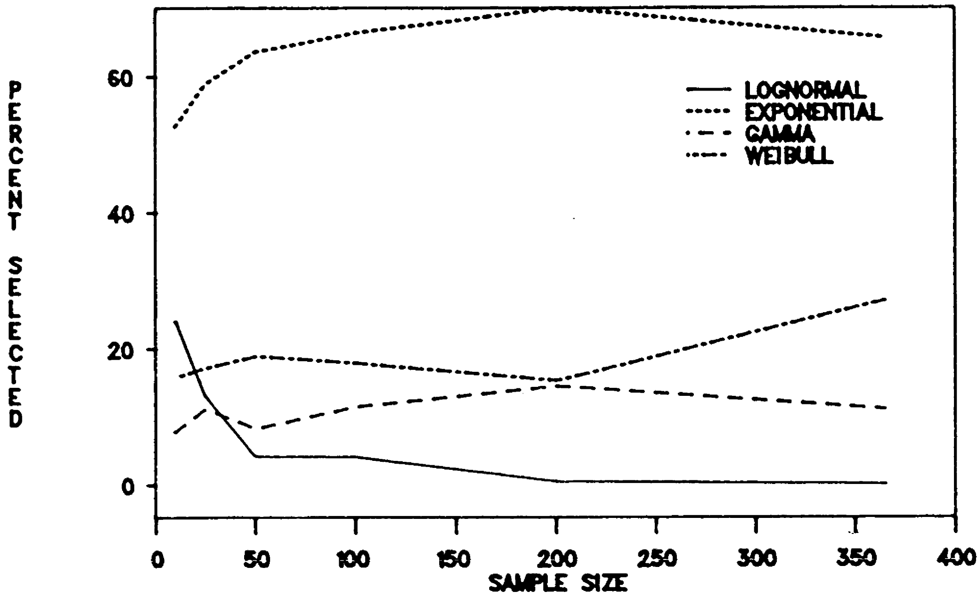

Figure 4.1: Probability of selecting a model using the Kolmogorov criteria with samples drawn from the exponential distribution.

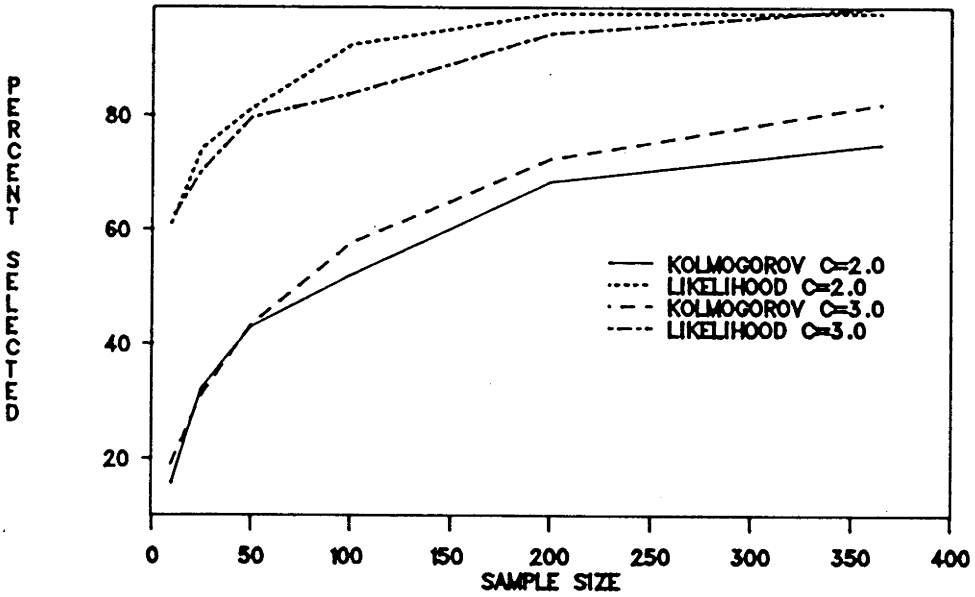

Figure 4.2: Probability of correctly selecting the gamma model for both the Kolmogorov and likelihood criterion.

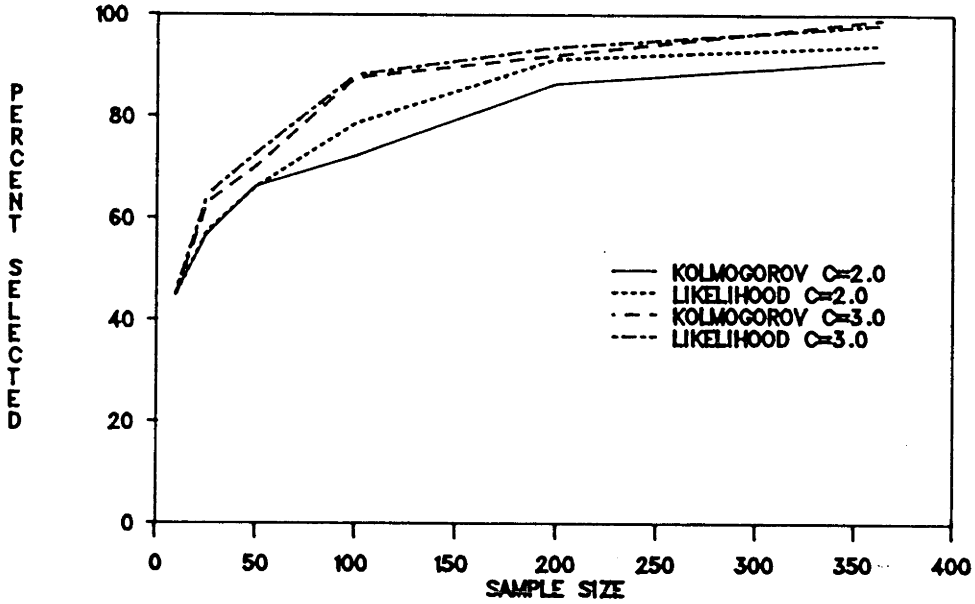

Figure 4.3: Probability of correctly selecting the Weibull model for both the likelihood and Kolmogorov criterion.

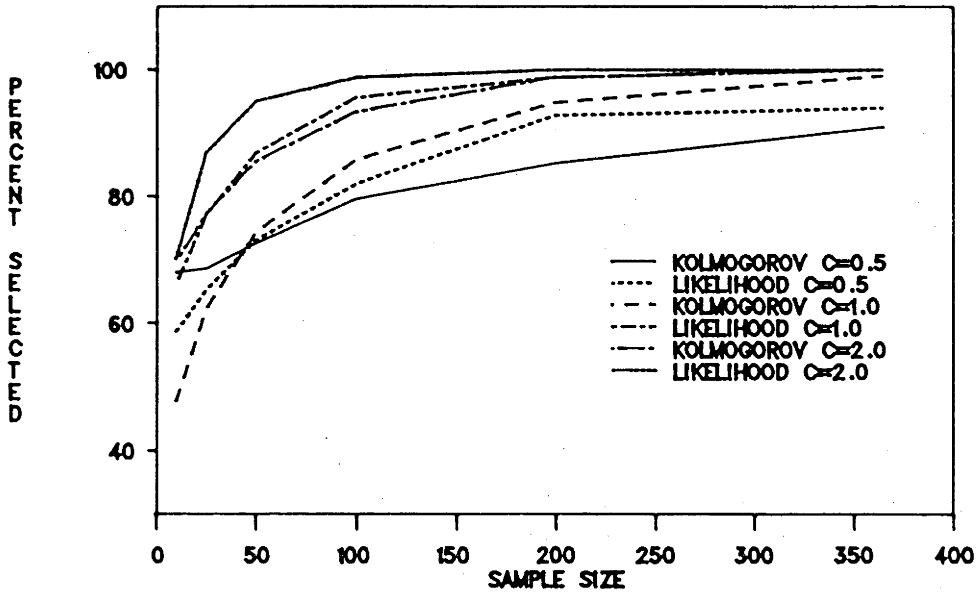

Figure 4.4: Probability of correctly selecting the lognormal model for both the likelihood and Kolmogorov criterion.

An explanation for the exponential distribution not being selected with an equal probability of approximately 0.33 with the gamma and Weibull models is because the maximum likelihood estimators for the parameters of both the gamma and Weibull distributions are postively biased. The implication being that the test statistic based on log likelihood is particularly sensitive to the method of estimation of the unknown parameters. Bain and Engelhardt (1980) noted that the test statistic for selecting between the Weibull and gamma models based on the ratio of the respective log likelihood functions was sensitive to the method of estimation of the unknown parameters. They recommended that simpler estimates should not be substituted for the maximum likelihood estimates. That the maximum likelihood estimators of the parameters of the gamma distribution are always positively biased has been shown by Berman (1981). For the Weibull distribution the positive bias has been observed over the range of sample sizes and parameter values under consideration (Thoman et al., 1969).

In all the Monte Carlo experiments with the exponential distribution the log likelihood selection criteria selected either the Weibull or gamma distribution. Table 4.8 presents the results of selecting between the gamma and exponential models and similarly for the Weibull and exponential models. Selection of the gamma model over the exponential model using the log likelihood criteria occurs for all simulations and sample sizes. For the Weibull versus exponential case the exponential distribution is selected when using only the log likelihood functions, with low probability. In both cases selecting the exponential distribution as the minimum ratio of the Kolmogorov test statistic to its 95% confidence interval occurs with a probability of ~ 0.7 over the range of sample sizes considered. Clearly this is an improvement over the log likelihood procedure. The problem in the case of the log likelihood procedure may be rectified to some extent through the use of unbiased estimators of the parameters. However, in this case the probability of selecting the exponential distribution will be equal to that of selecting any two-parameter model in which the exponential distribution is a special case. Thus where the gamma, Weibull and exponential models are considered the models will each be selected with a probability of 0.33. By comparison the Kolmogorov statistic yields a probability of correctly selecting the exponential distribution from amongst four possible distributions, including the gamma and Weibull models, above 0.5 for all the sample sizes considered here.

4.5.3 Model identification procedure

Clearly from the discussion of the simulation results no one method is likely to provide the 'best' selection criteria. The Kolmogorov based criteria yields an improved probability of correct selection when the true model is the exponential distribution. However where the choice is amongst the two-parameter lognormal, Weibull and gamma models the maximum value of the log likelihood functions will select the correct model with about equal to higher probability. Figure 4.2 illustrates that this difference in correct selection is most significant when the correct model is the gamma distribution.

Table 4.8: Probability of correctly selecting the exponential distribution according to log likelihood ratio and T_n/D_{95} tests, at various sample sizes.

Sample Size	Exponential		Gamma	
	T_n/D_{95}	Likelihood	T_n/D_{95}	Likelihood
10	72.7	0	27.3	100
25	74.3	0	25.7	100
50	72.0	0	28.0	100
100	70.4	0	28.6	100
200	74.4	0	25.6	100
365	70.0	0	30.0	100

Sample Size	Exponential		Weibull	
	T_n/D_{95}	Likelihood	T_n/D_{95}	Likelihood
10	69.1	13.9	30.9	86.1
25	69.2	3.4	30.8	96.6
50	65.6	0.8	34.4	99.2
100	69.2	0	30.8	100
200	73.2	0	26.8	100
365	66.0	0	34.0	100

When selecting a model appropriate for air quality data it will be necessary to have the facility to reject all models. This should allow new models to be developed when necessary. However where the maximized log likelihood is applied alone as the selection criteria, one model will always be selected.

It is proposed then, for selecting between the exponential, Weibull, gamma and lognormal models, that the Kolmogorov and log likelihood criteria be combined to yield an optimal selection criteria. In this case only the models accepted at a suitable confidence level, for example the 95% confidence level, are considered. At this stage all models may be identified as inapplicable. If amongst the models accepted the exponential model yields the minimum value of the Kolmogorov ratio then this model is accepted. If this is not the case then the log likelihood function is evaluated for each of the remaining models and the model with the maximum value selected. This procedure should provide an excellent method for selection among the exponential, Weibull, gamma and lognormal models while retaining the ability to reject all models.

4.6 Identification of a distributional model for air quality data recorded in Melbourne, Australia

Using the procedure for the investigation of goodness-of-fit of the exponential, gamma, lognormal and Weibull distributions to air quality data as discussed in the preceding section, the hypothesis was examined that all 24-h averaged air quality data are best described by a two-parameter lognormal distribution. The data examined are suspended particulates (β-scattering), nitric oxide, nitrogen dioxide, nitrogen oxides, ozone, sulphur dioxide and carbon monoxide recorded in Melbourne, Australia at 19 sites. Table 4.9 lists the monitoring sites with a brief description of the surrounding area and the period of operation. From Table 4.9 it may be readily ascertained that the monitoring sites are representative of the range of land use conditions to be found in and around a major urban area. The data sets span the entire period of operation of the monitoring network.

Table 4.9: A brief description of the monitoring sites in Melbourne, Australia

Monitoring station	Monitoring environment	Data available during years
Museum	Central business district	1975-1984
Alphington	Residential/light industrial	1978-1984
Maribyrnong	Industrial	1975
Monash	Light industrial	1975
Traralgon West	Residential/industrial	1975
Materials Research Laboratory	Industrial	1976
Watsonia	Residential	1976
Geelong	Industrial	1976-1978
Box Hill	Residential	1977
Flynn	Residential	1976-1977
Rosedale	Residential/industrial	1978
Morwell East	Residential/industrial	1978-1984
Parliament Place	Central business district	1975-1976 1978-1980
Mt Cottrell	Rural	1981-1984
Camberwell	Residential	1981-1984
Footscray	Industrial	1981-1984
Westmeadows	Semi-rural	1979-1984
Traralgon	Residential	1981-1984
Moe	Residential	1981-1984

The measurement of the pollutants is achieved using a variety of techniques (Environment Protection Authority of Victoria, 1983). For ozone and the nitrogen oxides chemiluminescence is used. Carbon monoxide concentrations are determined using infra red absorption while sulphur dioxide concentrations are measured using flame photometry with a hydrogen flame. A nephelometer measures the amount of light scattered by suspended particulates.

For each year of data, 24-h average concentrations were constructed from hourly averaged data. Only data sets with more than 100 24-h averages were considered in the analysis. Using the procedure discussed above for each data set the T_n/D_{95} statistic and log likelihood function were evaluated. The exponential distribution was selected when the T_n/D_{95} statistic was at a minimum, otherwise the distributional model with the largest value of the log likelihood function was chosen. The results of this analysis are presented in Figure 4.5. For each pollutant the frequencies with which each distributional model was selected from amongst the exponential, gamma, lognormal and Weibull models are given. The results clearly indicate that the two-parameter lognormal model, on the basis of the goodness-of-fit tests considered here, is not the best model for all pollutants. In fact the lognormal distribution was selected as the best distributional model for only 40% of all the data sets. However, for suspended particulates, measured as β-scattering, the lognormal distribution does appear to provide the best model for the observations with all but one data set selected as lognormal.

For the sulphur dioxide data sets the lognormal distribution was found to be representative of the majority of data sets. However a gamma model, and to a lesser extent the Weibull model, should be given consideration in any analysis of sulphur dioxide concentrations generated by area sources. It is noted that Berger et al.(1982) found that a two-parameter gamma model provided a better description than the usual two-parameter lognormal for the sulphur dioxide data sets examined in the Gent region of Belgium, although they also indicated that a two-parameter exponential form may be appropriate for the extreme concentrations.

For the carbon monoxide data sets studied it was observed that a gamma distribution was the model selected most often. The major alternative to the gamma model was the Weibull model. Clearly the gamma and Weibull models should be considered in any study of the distribution of carbon monoxide observations.

For ozone the distributional model selected for nearly all data sets was gamma. The lognormal and Weibull models were found to provide the best models for very few data sets. Similarly for nitrogen dioxide, the gamma model has found greatest acceptance. Trijonis (1978) also observed that the gamma model provided a better representation of nitrogen dioxide data than the lognormal.

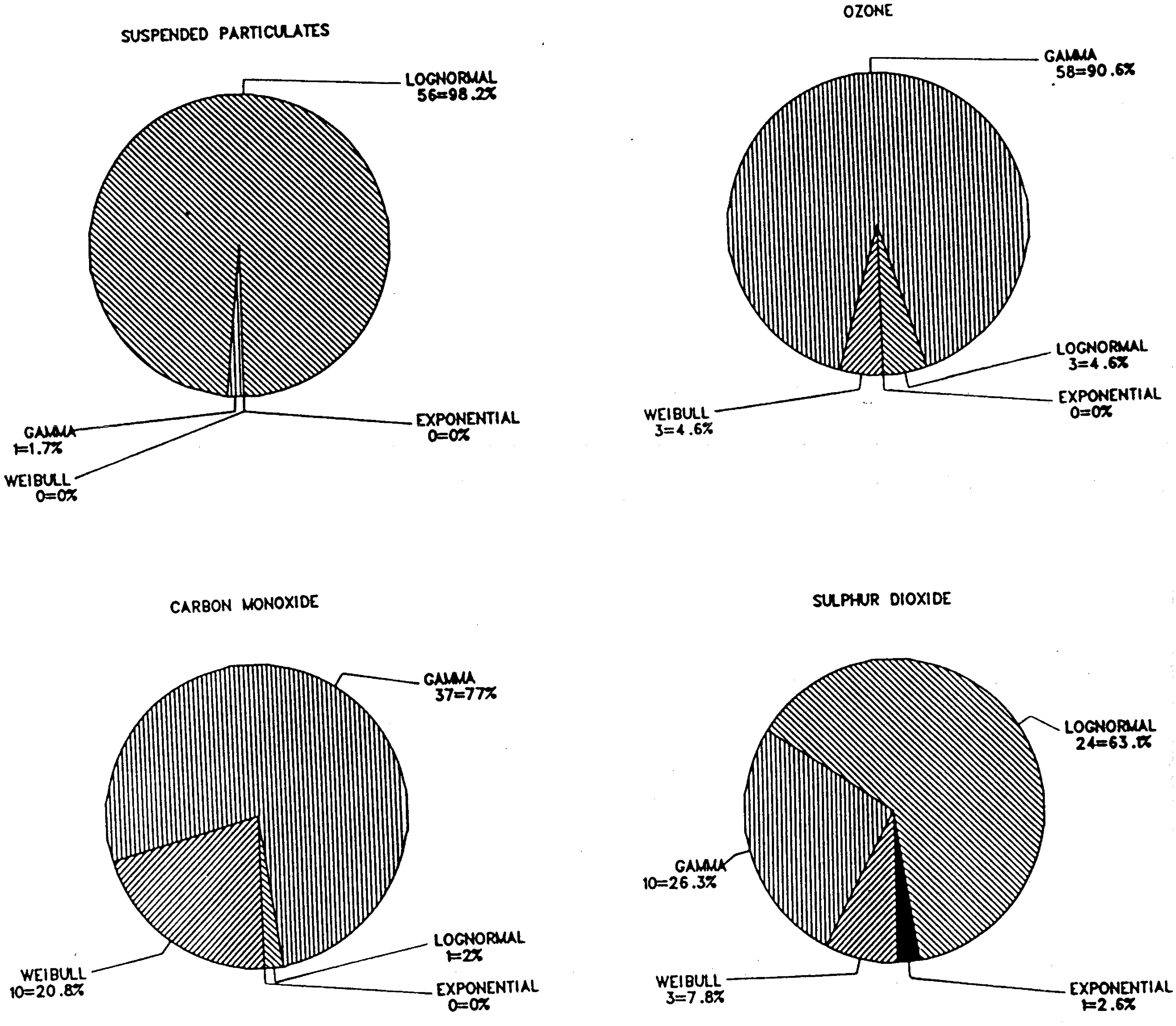

Figure 4.5: The frequency with which the exponential, gamma, lognormal and Weibull distributional models were selected for 24-h averaged air quality data recorded in Melbourne, Australia.

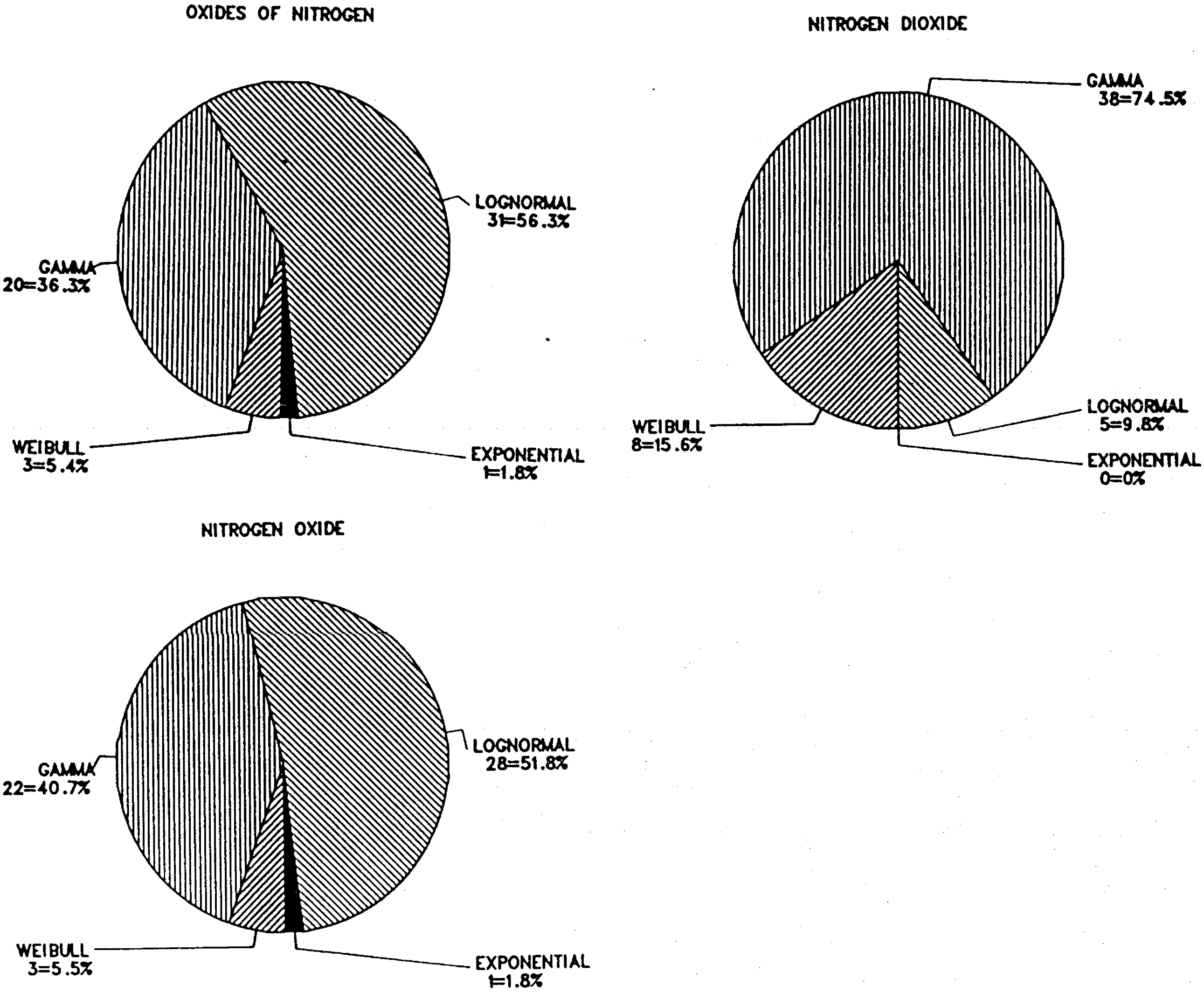

Figure 4.5 (Cont.)

This situation is reversed for nitric oxide and for oxides of nitrogen. Here the lognormal distribution is preferred but the gamma distribution remains an important model. That nitric oxide data should prefer a lognormal distribution while nitrogen dioxide is better described by a gamma distribution is related to the frequency of occurrence of the higher concentrations. Table 4.10 lists the yearly mean and maxima for nitrogen dioxide and nitric oxide recorded at two sites over several years. These data sets are considered representative of the entire data set. The data of Table 4.10 indicate that while the mean concentrations are about the same for the two pollutants substantial differences in the maximum concentrations were recorded. It is this tendency towards a long tailed distribution which produces the preference for the lognormal distribution in the case of nitric oxide. Oxides of nitrogen represent the total of the nitric oxide and nitrogen dioxide observations. It would appear that the greater frequency of observations in the upper percentiles of the nitric oxide distribution produces the observed preference of the lognormal model for oxides of nitrogen.

Table 4.10: The yearly mean and maximum concentrations (ppm) of nitrogen dioxide and nitrogen oxide recorded at the Alphington and Museum monitoring stations in Melbourne, Australia 1978-1981.

Monitoring site	Year	Nitrogen dioxide		Nitric oxide	
		mean	maximum	mean	maximum
Alphington	1978	0.016	0.041	0.007	0.030
	1979	0.012	0.040	0.018	0.140
	1980	0.013	0.040	0.019	0.170
	1981	0.013	0.030	0.016	0.120
Museum	1978	0.025	0.077	0.060	0.344
	1979	0.020	0.050	0.041	0.200
	1980	0.025	0.070	0.040	0.150
	1981	0.021	0.060	0.031	0.160

4.7 Conclusions

When identifying normal, or by transformation lognormal, distributions for air quality data it was found that Lilliefors test and the D'Agostino statistic were more powerful than the chi-square test for the wide range of non-normal populations considered. Overall the Lilliefors test was the more powerful for air quality data. A selection procedure which combined the Lilliefors statistic with the log likelihood function to identify the best model from amongst the two-parameter lognormal, two-parameter gamma, two-parameter Weibull and one-parameter exponential alternatives was developed. Then the problem of selecting the best distributional form for an extensive air quality data set recorded in Melbourne, Australia was considered. It was found that the two-parameter lognormal distribution was best for suspended particulates and for the majority of the nitric oxide, oxides of nitrogen, and sulphur dioxide data sets. The gamma distribution was found to provide the best distributional model for ozone, nitrogen dioxide and carbon monoxide observations, while the Weibull form was significant for carbon monoxide and sulphur dioxide data.

CHAPTER 5
A MODEL FOR PREDICTING THE DISTRIBUTION OF AREA SOURCE ACID GAS CONCENTRATIONS

5.1 Introduction

A hybrid modelling approach was recently applied by Simpson et al. (1983) to predict the maximum 24-h average pollutant concentration for carbon monoxide in Canberra, Australia and total suspended particulates in Brisbane, Australia. Simpson and Jakeman (1984) invoked the same model to predict maximum 24-h average acid gas concentrations for Newcastle, Australia. They showed that over a 10 year period a variation in annual maxima of the order of a factor of 2 is to be expected simply due to changes in the windspeed distribution.

In this chapter, the hybrid modelling approach is applied to predict the upper percentiles, and in particular the 98-percentile, of 24-h average acid gas data collected at 6 sites within the urban area of Newcastle, Australia over 8-10 years. Estimates of the 98-percentile concentration are evaluated for direct comparison with the World Health Organisation goal for acid gases which is $200\ \mu\ gm^{-3}$, a level which is not to be exceeded more than 2% of the time.

The model applied here employs the same deterministic component as that of Simpson and Jakeman (1984) in their study of acid gas concentrations from two of the monitoring sites in Newcastle. However, for the 55 records of yearly 24-h average data available, the gamma distribution is identified as a more appropriate statistical component than the previous lognormal distribution for the prediction of the 98-percentile concentration. The work illustrates the importance of powerful methods of identification for determining the most appropriate distributional form of pollution data. Identification methods based upon the chi-square test accept the data analysed in Simpson and Jakeman (1984) as lognormal but the more powerful test developed in Chapter 4 of this thesis prefers a gamma representation over Weibull, exponential and lognormal types. As we shall see, this leads to increased flexibility in application of the model.

In this chapter a method to quantify the level of uncertainty associated with the hybrid model predictions is presented. This is achieved by applying the maximum likelihood method to fit the truncated distribution of pollutant concentration produced by the deterministic model component. This yields parameter estimates with normally distributed statistical properties. A Monte Carlo simulation exercise is then performed to yield approximate 95% confidence levels for the percentiles of the distribution. This theoretical evaluation of the confidence levels, at least for the data sets examined in this study, provides realistic estimates of the uncertainty in model predictions.

5.2 The area source hybrid model

The model of Simpson and Jakeman (1984) assumes that : (1) the air pollution data and wind speed data are lognormally distributed and (2) there is, on average, an inverse relationship between wind speed and air pollution levels. Larsen (1969, 1971) has observed using graphical methods that condition (1) is valid for many air pollution data sets. This was also observed for those data sets used in the Australian examples referred to above by plotting them on lognormal-probability axes (see for example, Simpson and Jakeman, 1984, Fig. 1). Condition (2) can be derived from the ATDL model developed by Gifford and Hanna (1973) where concentration χ is related to wind speed u according to

$$\chi = CQ/u = K'/u \qquad (5.1)$$

where K'=CQ is a constant dependent on source strength Q, and atmospheric stability, C. Simpson et al. (1983) use instead the following percentile relationship, an approach also used by Benarie (1976) and Knox and Lange(1974)

$$\chi_p = K/u_{100-p} \qquad (5.2)$$

where x_p is the p-percentile of concentration and u_{100-p} is the opposite percentile of the distribution of wind speed. Simpson and Jakeman (1984) show that K is reasonably constant for the $30<p<70$ percentiles for their acid gas data. The equation (5.2) differs from equation (5.1) in that it only requires opposing percentiles to be related by a constant rather than time-wise pairs of pollution and windspeed.

The hybrid procedure based upon equation (5.2) works as follows. The sample probability distributions of concentration and windspeed are constructed so that the range of values of p for which K is constant can be evaluated using equation (5.2). For a given windspeed distribution this permits those concentrations, over the percentile range for which K is constant, to be determined. This is the deterministic component of the hybrid model. It allows the prediction of pollutant concentrations but only within the range of reliability of the deterministic component as ascertained by analysis of actual data. The statistical component which is a phenomenological model (Benarie, ~~1974~~ 1982) in that it is non-causal or non-predictive, containing the observed history of the pollution phenomena, is then invoked in a two step procedure. Firstly, the relevant parametric form of the distribution of pollution data (lognormal in the previous examples) is identified using the statistical tests developed in Chapter 4. Secondly, the parameters of the identified form are estimated (for example, by least squares curve fitting or by applying the method of maximum likelihood) using those percentile points reliably predicted from the deterministic model. Once the parameters are known, the full distribution including the maximum can be inferred.

The hybrid model is not limited by the assumption of lognormality. As long as assumption (5.2) is valid in the sense that within a specified range the opposing percentiles of wind speed and air pollution are related by a constant, any identified distributional form of the pollution data can be used. Equation (5.2) is quite general and does not depend on lognormality of the pollution nor the wind speed data.

In order to identify an appropriate model for the distribution of the acid gas data sets, goodness-of-fit tests of the Kolmogorov type and the maximum of the log likelihood function were employed to compare the two-parameter lognormal, gamma and Weibull distributions and the exponential distribution. The parameters of these distributions were estimated using the method of maximum likelihood. The results of the goodness-of-fit tests to the 55 data sets indicated that the gamma distribution was preferred over the lognormal for 41 of the data sets. Such a result indicated that a gamma model should be a better representation of the whole ensemble than the lognormal model. However this result does not necessarily imply that the gamma distribution will provide the best model for estimating maximum concentrations.

5.3 The data set and model assumptions

Newcastle is recognised as a major industrial city and seaport. The industrial area consists of such industries as steelworks, chemical and metallurgical plants and brickworks and are located along the Hunter River with the major population areas to the south. Data are recorded at 6 sites, referred to here as the Watt Street, Mounter Street, City Hall, Turton Road, Elder Street and Seaview monitors, and all have been employed in this study (see Figure 5.1). The 3 sites, the Watt Street, Mounter Street and City monitors lie in close proximity to the industrial area and have records over the 10 year period, 1972-1981. The remaining three sites lie within 5-10 km of the industrial complex in the residential area of Newcastle and recordings are available for the years 1973-1981 for Turton Road and for the years 1974-1981 for both the Elder and Seaview monitoring locations.

The acid gas data set used consists of observations of 24-h average acid gas levels measured by the Health Division of the Newcastle City Council at 6 sites within the city of Newcastle, NSW, Australia. The acid gas concentrations were determined using the British Standard Method No. 1747 Part 3. The acid gas levels were obtained for the 24 hour period beginning at 9am for 5 days per week extending over the full year.

The windspeed data were recorded as 10 minute averages using a Dines anemometer every 3 hours at the Williamtown Airport weather station which lies about 20-25km to the north of the air quality monitoring stations. With these data average windspeed measurements were computed for the 24 hour period from 9am each day.

It should be noted that the model developed here is applicable not only to the 2 sites originally examined by Simpson and Jakeman (1984), but to all sites at which measurements are available. This result was obtained by relaxing the assumption of Simpson and Jakeman (1984) that both the air pollution data and windspeed data are lognormally distributed. In particular the close agreement between the geometric standard deviations of the windspeed and pollutant data sets is not required. This correlation is a requirement for the estimation of the maximum pollutant concentration (x_{max}) using the lognormal assumption. In the latter case the maximum is calculated according to the expression

$$x_{max} = \frac{K}{\alpha_u} \beta_u^{Z_m} \tag{5.3}$$

where β_u is the geometric standard deviation of the windspeed data set, α_u the geometric mean and Z_m the number of standard deviations from the mean corresponding to the percentile point for the maximum value. If β_u in equation (5.3) does not closely correspond with that evaluated from the pollutant data, significant errors (greater than a factor of 2) may arise.

For all windspeed and pollutant data sets it is required that the condition as stated by equation (5.2) holds, that is, there is on average, an inverse relationship between windspeed and pollutant concentration. Simpson and Jakeman (1984) have clearly demonstrated that this condition is met for the observations recorded at the Watt and Mounter sites. Table 5.1 presents a representative sample of K values obtained for the 30-70 percentiles using the method of Simpson and Jakeman (1984) for 2 years of data for each of the City, Turton, Elder and Seaview monitors. Figure 5.2 illustrates the variation of the K-value over all sites for the data sets of 1978 while Figure 5.3 presents the K-values obtained for the Elder monitor. Figures 5.2 and 5.3 and the data listed in Table 5.1 are considered representative of the variation of K

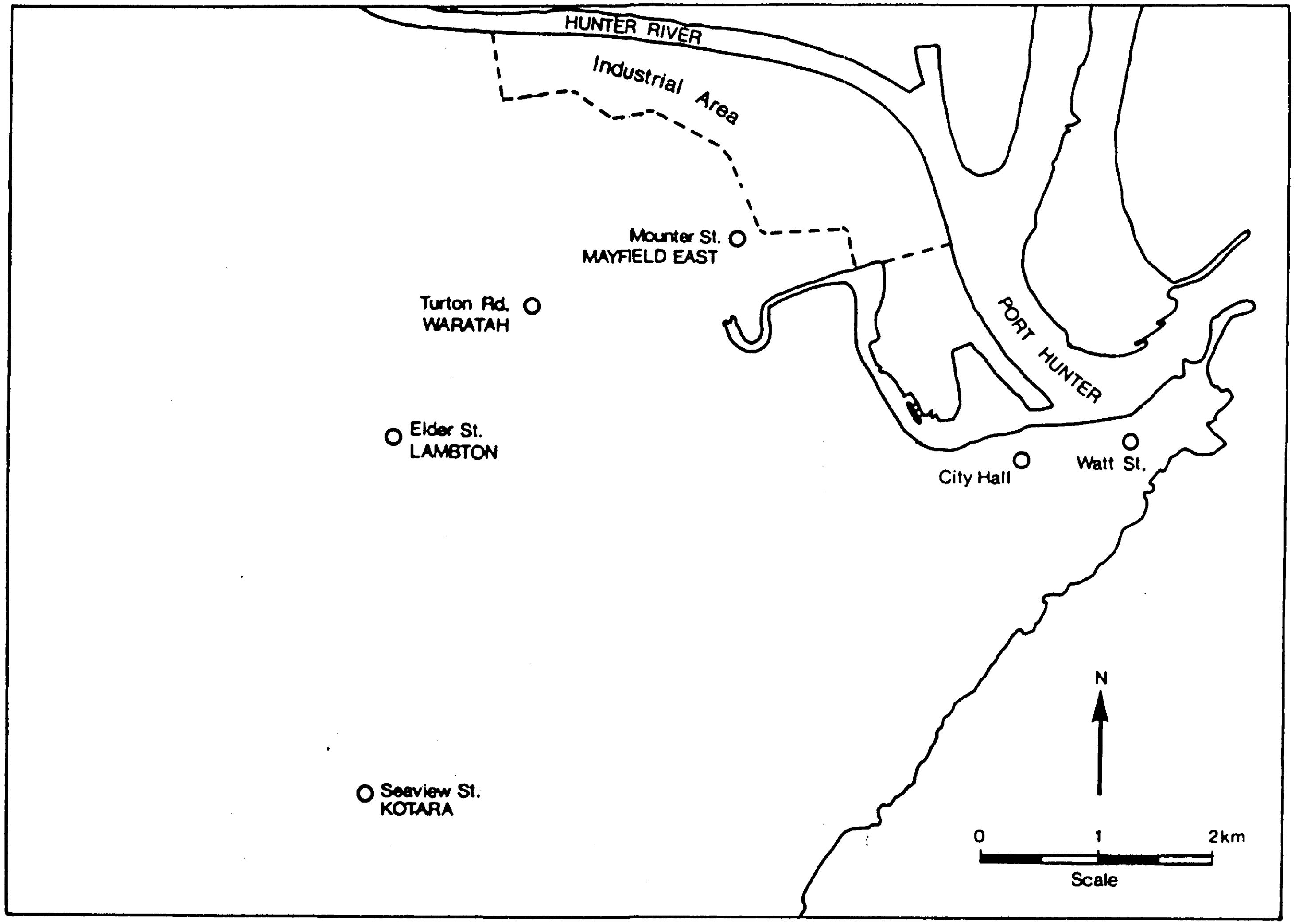

Figure 5.1: Newcastle, Australia, acid gas monitoring network.

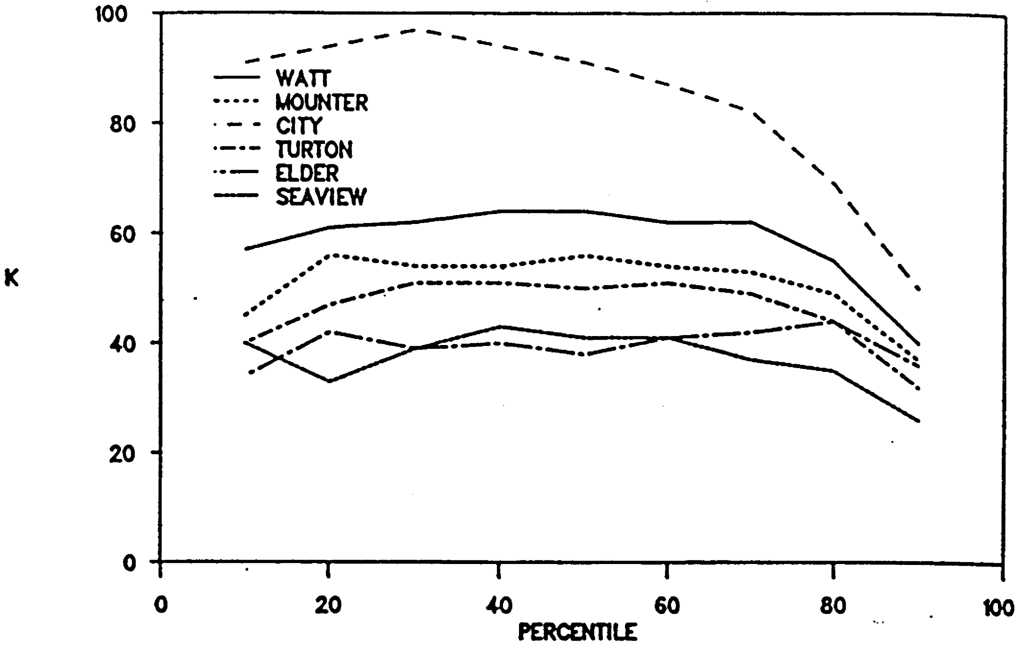

Figure 5.2: Variation in K ($\mu gm^{-2}s^{-1}$) for all sites for the year 1978.

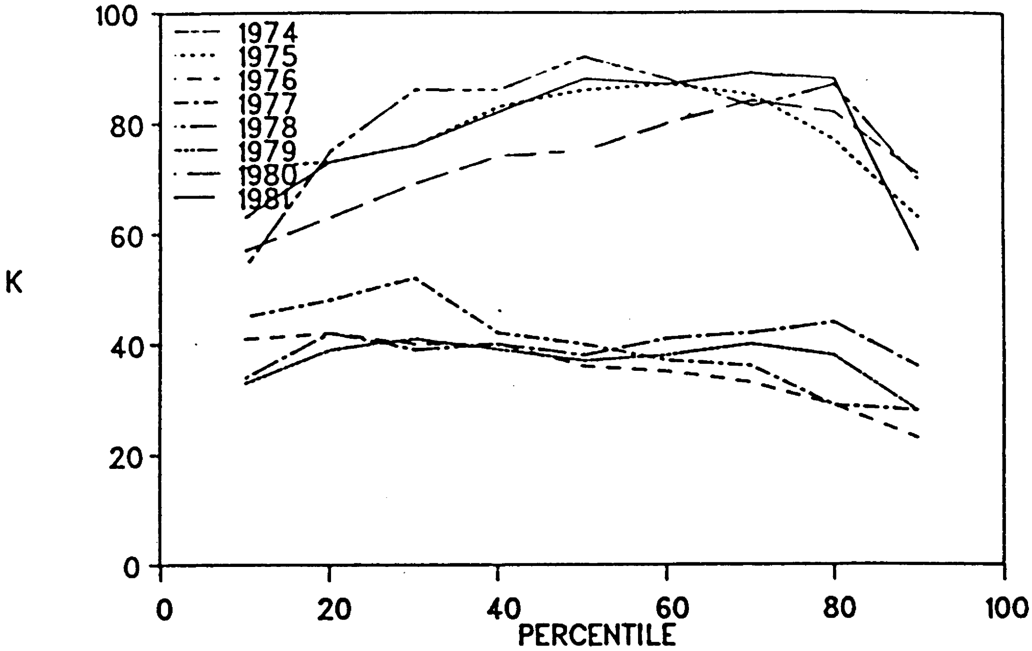

Figure 5.3: K ($\mu gm^{-2}s^{-1}$) determined at the Elder Street monitor for all years.

observed over the whole sample of acid gas concentrations. These values indicate that equation (5.2), except in the case of the City monitor, is a good approximation over this range of percentiles. As shall be demonstrated, the results of the hybrid model prediction of the 98-percentile are poorest for the City site.

Table 5.1: Estimates of K $(\mu g\ m^{-2}\ s^{-1})$ for a range of deciles.

Year	Site			K at decile		
		30	40	50	60	70
1973	City	48	46	45	49	52
1974	Turton	139	144	144	144	152
1975	Elder	76	82	86	87	85
1976	Seaview	36	40	36	35	29
1977	City	95	99	95	94	85
1978	Turton	47	46	46	44	44
1979	Elder	41	39	38	38	40
1980	Seaview	70	74	75	79	80

The debate concerning the simple relationship, described by equation (5.2), is examined in detail by Simpson et al. (1985). The interpretation of this simple relationship does not form part of this thesis. Suffice it to say here that the deterministic component of the hybrid model, equation (5.2), does provide reliable estimates of percentiles in the range $30<p<70$ as evidenced empirically by the constancy of the K-factors (see Table 5.1) over all sites. The effect of variation in the K-factor will be examined in detail in section 5.5. Thus assuming that the data ~~satisfies~~ satisfy the conditions for the application of the model, estimation of the 98-percentile concentration will now be considered.

5.4 Estimating the 98-percentile pollutant concentration

For the two-parameter gamma distribution the 98-percentile (x_{98}) may be obtained as the solution to the equation

$$0.98 = \frac{1}{\Gamma(c)} \int_{o}^{x_{98}} e^{-t/b} (t/b)^{c-1} dt \qquad (5.4)$$

where Γ is the gamma function, b the scale parameter and c the shape parameter. Equation (5.4) can be solved numerically given estimates of the scale and shape parameters.

The estimation of the parameters of the gamma distribution from the truncated sample (i.e. for the restricted percentile range $30 < p \leq 70$) was performed using the method of maximum likelihood. This approach selects parameters of the gamma distribution in which the observed concentrations occur with the highest joint probability. These estimates are the values of the gamma scale parameter b, and shape parameter c, which maximize the likelihood function. When the first r smallest values and last n-m largest values are censored yielding the (m-r) order statistics $x_{r+1}, x_{r+2}, \ldots, x_m$ from a total sample of size n the natural logarithm of the likelihood function is known as (Cohen and Norgaard, 1977)

$$\ln L = -(m-r) \ln \Gamma(c) \quad -(m-r)c \ln b$$

$$-\frac{1}{b} \sum_{i=r+1}^{m} x_i + (c-1) \sum_{i=r+1}^{m} \ln (x_i) \qquad (5.5)$$

$$+ (n-m) \ln \{1 - F(x_m)\} + r \ln \{F(x_{r+1})\}$$

where $F(x)$ is the cumulative gamma distribution function. Using the partial derivatives with respect to the shape and scale parameters an iterative search algorithm was applied to determine the maximum value of the likelihood function. The inverse of the information matrix, the variance-covariance matrix, was evaluated for each data set based on the maximum likelihood parameter estimates.

In order to determine the confidence intervals for a given percentile a simple Monte Carlo procedure was utilised in conjunction with equation (5.4). 200 samples of each of the gamma distribution parameters were selected from a correlated multivariate normal distribution (Naylor et al., 1966) with the mean values being the maximum likelihood estimates of the gamma parameters, and the standard deviations being derived from the variance-covariance matrix associated with the maximum likelihood estimates. For each pair of parameters the 98-percentile was calculated according to equation (5.4). The 200 estimates of the 98-percentile were ordered and the approximate 95% confidence levels determined. The choice of a sample size of 200 represents a balance between excessive computing time and accuracy. The sample size of 200 provides estimates of the confidence bounds to within a few percent, which is sufficiently accurate for the purposes of obtaining approximate confidence bounds and demonstrating their usefulness.

Table 5.2: Estimates of K (μg m^{-2} s^{-1}) for all sites

Year	Watt	Mounter	City	Turton	Elder	Seaview
1972	98	98	59	-	-	-
1973	61	51	45	74	-	-
1974	76	97	71	144	88	85
1975	73	82	80	132	86	97
1976	54	62	36	46	36	36
1977	55	49	95	46	41	38
1978	61	52	91	46	38	41
1979	57	68	65	52	38	52
1980	90	87	95	95	75	75
1981	92	67	80	45	92	87

The K-factors for the Watt and Mounter data are those derived by Simpson and Jakeman (1984). These were calculated simply as the product of the medians of the distribution of pollutant concentration and windspeed. The K-factors for the City, Turton, Elder and Seaview monitors were obtained in the same way. Table 5.2 presents those K-factors. For the Watt and Mounter sites, the more rigid assumptions of the Simpson and Jakeman (1984) model also apply since the lognormal assumption of pollution concentration and windspeed are reasonable for these data sets. Hence we can calculate estimates of the 98-percentile concentration derived from the application of the model of Simpson and Jakeman (1984) and compare them with those of the model developed in this paper. In the former case, the p-percentile concentration, x_p, can be obtained from

$$x_p = \frac{K}{\alpha_u} \beta_u^{Z_p} \qquad (5.6)$$

where α_u is the median windspeed and β_u is the geometric standard deviation, Z_p is the number of standard deviations from the mean to the p-percentile and K is as listed in Table 5.2. Using values of α_u and β_u obtained by Simpson and Jakeman (1984), and a value of Z_p of 2.054 which corresponds to the 98-percentile, estimates of x_{98} were determined.

Table 5.3 presents the estimates of x_{98} obtained from both models. For the lognormal model the root mean square error between estimated and observed is 21.6 μg m^{-3} while for a gamma distribution this value is 24.1 μg m^{-3} . This result indicates that the gamma distribution yields estimates of the 98-percentile with similar accuracy to that of the lognormal model. However for the 1981 Watt and 1980 Mounter data sets the goodness-of-fit tests identified the lognormal distribution as more appropriate than the gamma distribution and this is reflected in the results of Table 5.3. If the estimates for these sites are omitted, the root mean square error between estimated and observed becomes 16.4 μg m^{-3} for the gamma distribution, which is lower than the 17.2 μg m^{-3} for the lognormal case. Figures 5.4 and 5.5 show the observed and predicted 98-percentile concentrations for the Watt and Mounter data sets using the hybrid model developed here.

Table 5.3: The estimated 98-percentile acid gas concentration ($\mu g\ m^{-3}$) for the Watt and Mounter monitoring sites using the lognormal and gamma models compared with the observed 98-percentile value.

	Watt			Mounter		
Year	Gamma Model	Lognormal Model	X_{98} observed	Gamma Model	Lognormal Model	X_{98} observed
1972	150	140	96	154	137	92
1973	44	51	52	53	43	46
1974	65	55	69	68	70	73
1975	72	87	69	84	98	83
1976	64	63	45	82	72	65
1977	67	88	42	65	78	44
1978	57	62	45	50	53	46
1979	67	58	60	77	69	62
1980	91	92	88	65	89	104
1981	115	136	146	85	95	81

For the City, Turton Street, Elder Street and Seaview monitoring sites the results are given in Table 5.4 and Figures 5.6 to 5.9. For these sites no comparison may be made with the model of Simpson and Jakeman (1984) as the assumptions on which that model were based are not valid for the data monitored there. In general, it can be seen that the model predicts the 98-percentile acid gas concentration to well within a factor of 2 in most cases, and that the observations generally fall within the predicted confidence intervals. Where the model does not provide close agreement with the observed 98-percentile the differences can mainly be attributed to the gamma distribution being inapplicable. In a minority of cases the simple inverse relationship of equation (5.2) is inappropriate even though the observations follow a gamma distribution. Consider the following cases where poor estimates of the 98-percentile were obtained. The 1981 Watt Street and 1980 Mounter Street data sets have already been mentioned where the Kolmogorov test preferred the lognormal distribution. Figures 5.2 and 5.3 show, for the 1978 City and 1980 Elder Street data sets, that the K-value does not remain constant indicating that equation (5.2) may not be valid. Similarly, for the 1972 Watt Street and Mounter Street data sets, the K-value decreases by over 30% from the 50 to the 70 percentile.

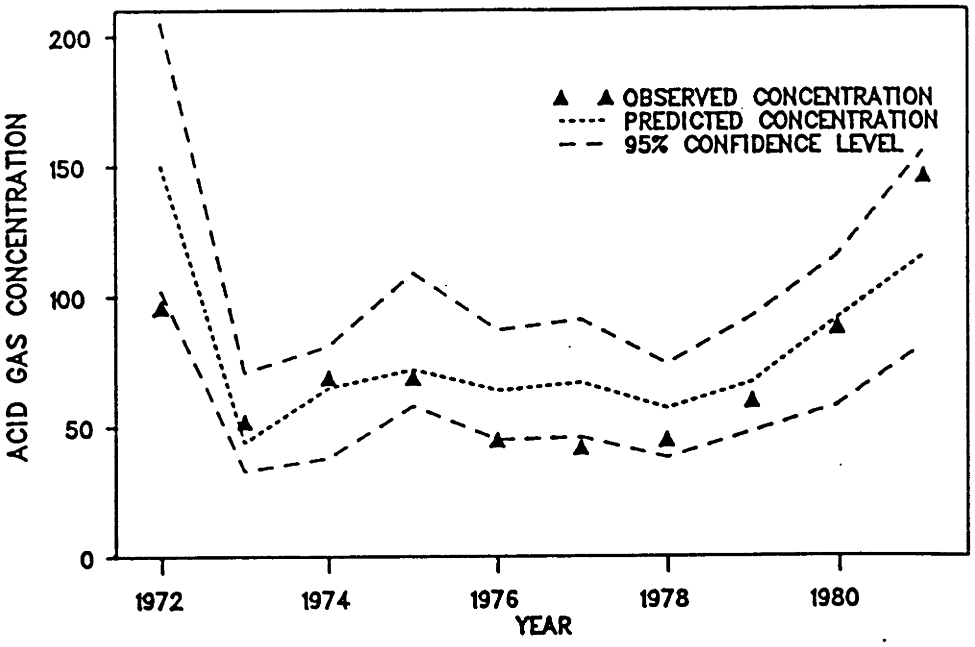

Figure 5.4: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the Watt Street data. The estimated 95% confidence intervals are also presented (- -).

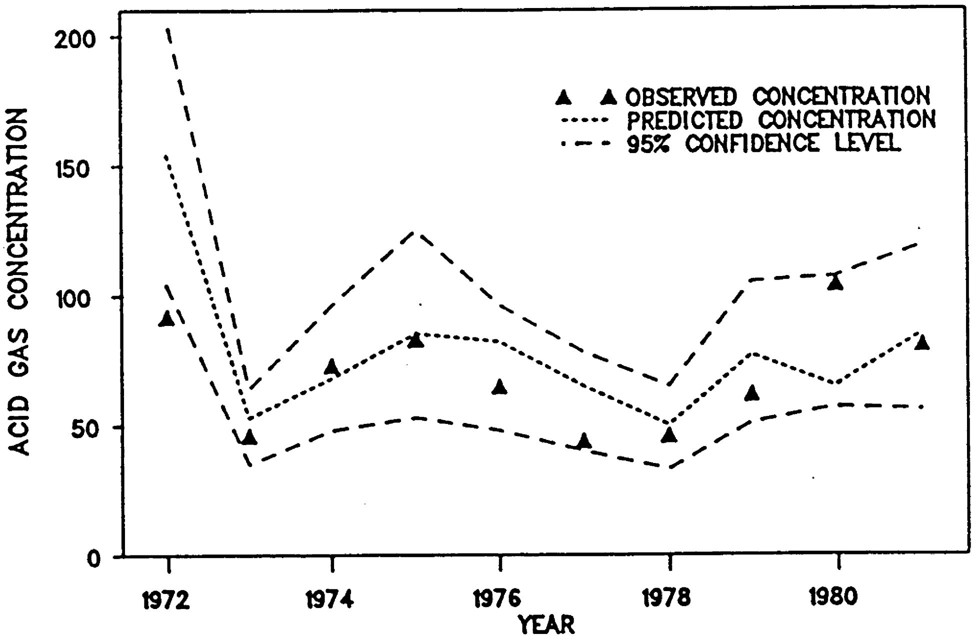

Figure 5.5: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the Mounter Street data. The estimated 95% confidence intervals are also presented (- -).

Table 5.4: Comparison of the estimated 98-percentile (x_o) with the observed 98-percentile concentration (x_{98}) for acid gas data ($\mu g\ m^{-3}$) at the City, Turton Road, Elder Street and Seaview monitoring sites.

Year	City		Turton		Elder		Seaview	
	x_{98}	x_o	x_{98}	x_o	x_{98}	x_o	x_{98}	x_o
1972	71	80	-	-	-	-	-	-
1973	42	42	124	62	-	-	-	-
1974	93	56	89	115	63	55	60	53
1975	94	83	140	145	95	84	94	100
1976	52	40	63	64	59	48	56	41
1977	76	111	38	54	43	50	26	45
1978	59	78	37	45	42	27	30	35
1979	54	71	56	57	40	41	42	64
1980	90	95	90	99	108	64	86	77
1981	121	88	70	50	126	102	116	106

5.5 Sensitivity of the 98-percentile estimate to the K-factor

For any given set of observations which are considered to follow a gamma distribution, the scale parameter b, may be estimated by

$$b = \frac{M_{1,c}}{M_{b,c}} \qquad (5.7)$$

where $M_{b,c}$ is the median of observations from a distribution with scale and shape parameters, b and c respectively, and $M_{1,c}$ is the median when the scale parameter is unity. In fact at any percentile p, by definition

$$b = \frac{x_{p,1,c}}{x_{p,b,c}} \qquad (5.8)$$

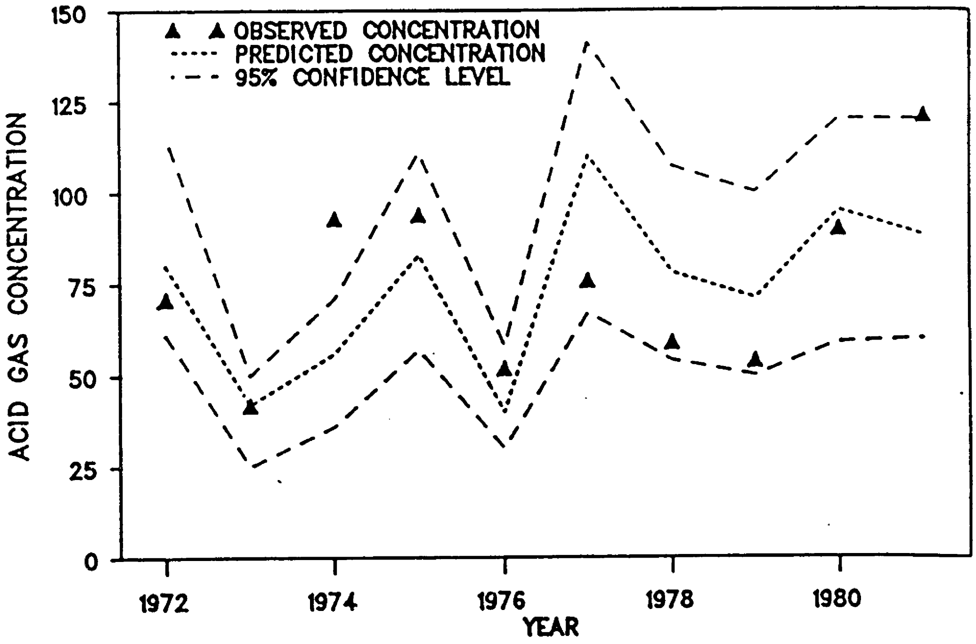

Figure 5.6: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the City Hall data. The estimated 95% confidence intervals are also presented (- -).

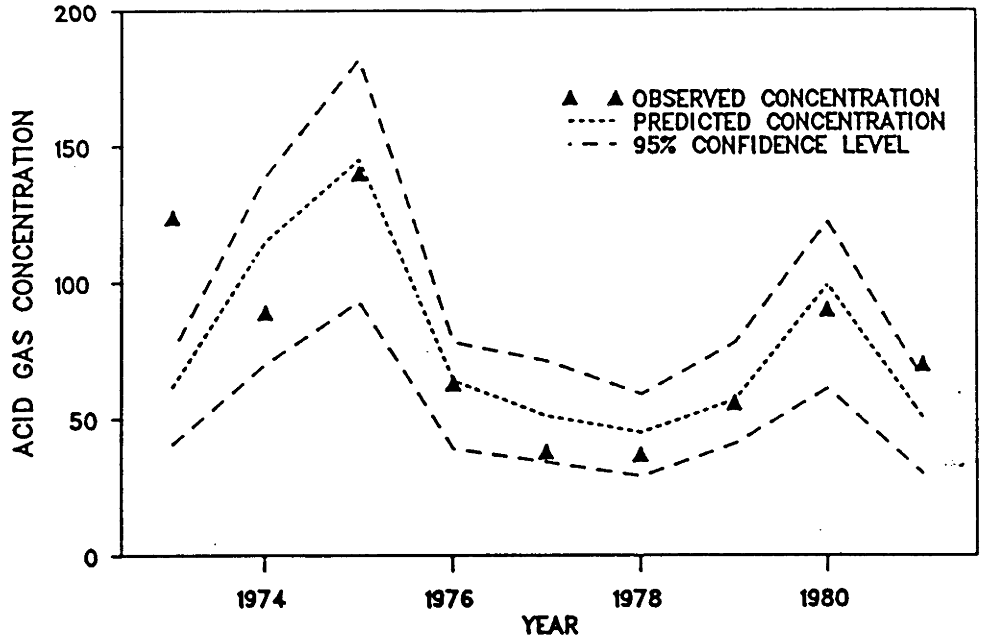

Figure 5.7: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the Turton Road data. The estimated 95% confidence intervals are also presented (- -).

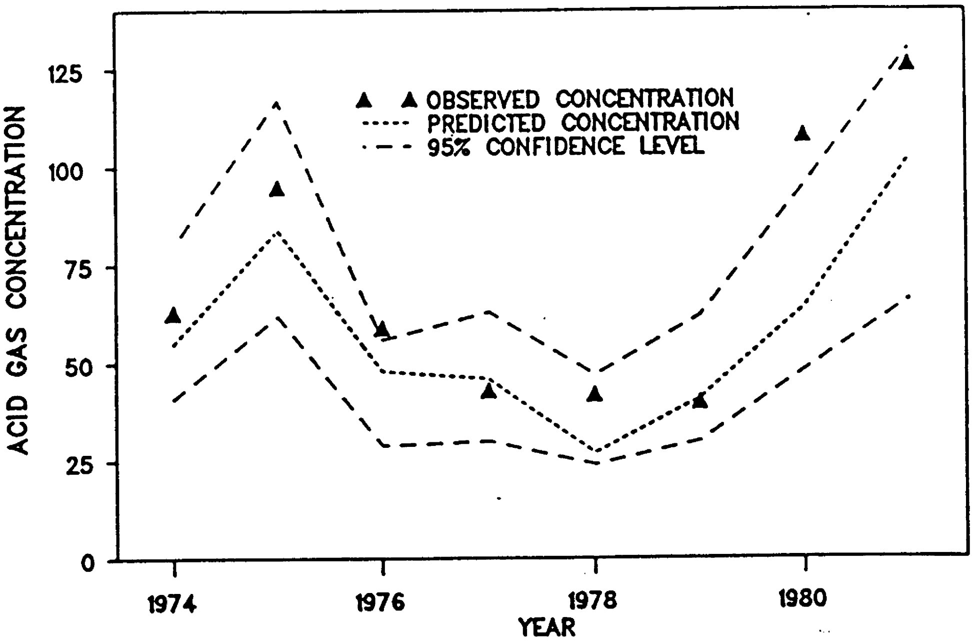

Figure 5.8: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the Elder Street data. The estimated 95% confidence intervals are also presented (- -).

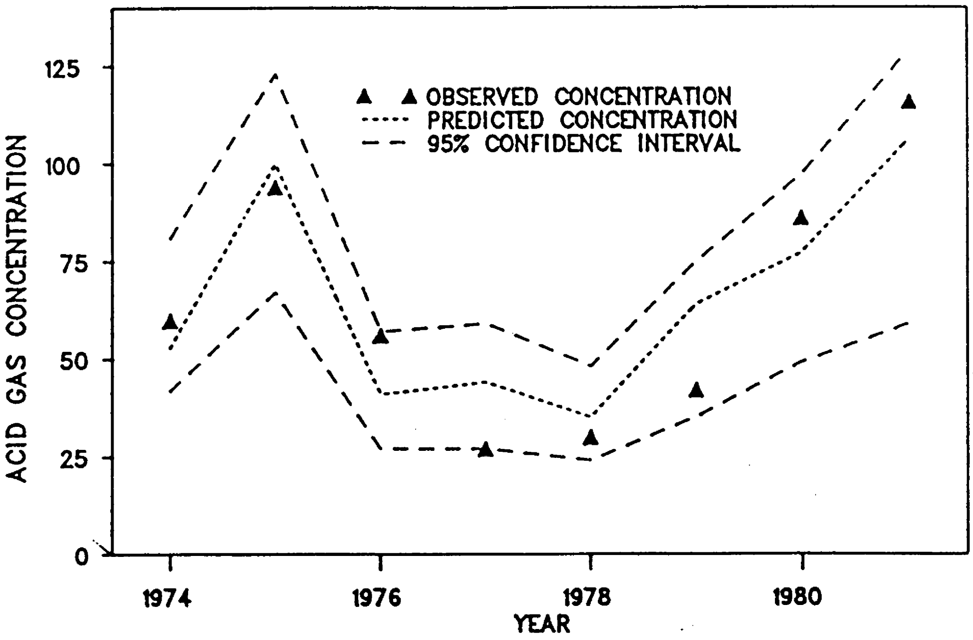

Figure 5.9: The observed and predicted (--) acid gas concentrations (μgm^{-3}) for the Seaview data. The estimated 95% confidence intervals are also presented (- -).

Now from equation (5.2) the K-factor of the deterministic model is related to the scale parameter as

$$M_{b,c} = \frac{K}{u_{50}} \tag{5.9}$$

where u_{50} is the 50-percentile or median windspeed value. Substituting equation (5.9) into equation (5.7) and rearranging obtains

$$b = \frac{M_{1,c}\ u_{50}}{K} \tag{5.10}$$

Thus, the K-factor is inversely proportional to the scale parameter, so that a given percentage uncertainty in the K-factor, will produce a percentage uncertainty in the scale parameter which may be determined by equation (5.10). As b is the scale parameter, equation (5.8) implies that this same percentage uncertainty due to errors in K will result for estimates of any percentile in the gamma distribution.

Estimates of the value of the K-factor evaluated over the entire 30-70 percentile range for all 55 data sets have an average relative standard deviation of approximately 5.0%. While it might appear from Figures 5.4-5.9 that the contribution to error in the estimate of the 98-percentile due to uncertainty in the K-factor is minor in comparison with the total uncertainty, the assumption of the constancy of the K-factor over the 30-70 percentile range has been examined in detail.

As a first step, the K-factor was evaluated for each percentile using the acid gas and windspeed data over the 30-70 percentile range by rearranging equation (5.2). These estimates were plotted against the order statistic corresponding to the acid gas data set. The hypothesis that the slope of a line through these data was non-zero was examined using the F-test for each of the 55 data sets. The hypothesis was rejected at the 95% confidence level on only 9 occasions. This result indicates that the K-factor varies systematically over the 30-70 percentile range although in most cases the slope, while significant, is close to zero. However the F-test is not strictly applicable to data which has been ordered as the test will tend to accept the hypothesis of a non-zero slope more frequently than the confidence level would suggest. Using the estimates of the slope derived previously, the hypothesis that these values were normally distributed was examined using the modified Kolmogorov test developed by Lilliefors (1967). The ratio of the test statistic to the 95% confidence level was 0.564 indicating that the hypothesis of normality may be accepted at this confidence level.

Based upon the assumption of normality the hypothesis that the slopes were drawn from a population with mean zero is considered next. The following test statistic was evaluated

$$z = \frac{u\, n^{1/2}}{s} \qquad (5.11)$$

where u is the sample mean, s^2 is the sample variance and n the sample size. A value of z = 0.595 resulted which may be compared with the 95% confidence level of 1.96. This means that the hypothesis that these data have mean zero can be accepted at the 95% confidence level. The result implies, when considering the 55 acid gas data sets, that there is no systematic deviation from the assumption of a constant value of K at least over the 30-70 percentile range.

In order to assess the likely effect of the variation in K for individual data sets a Monte Carlo study was performed. Here the value of K was allowed to vary over the percentile range employed to estimate the parameters of the gamma distribution. The value of K was assumed to vary linearly over this range. A suitable range of values for the slope, α_k , was determined by evaluating α_k for each of the 55 acid gas data sets. The range was found to be from -0.0015 to 0.0015. Differences in α_k resulting from the different magnitudes of K were eliminated by normalizing the K-values to the mean K-value as determined over the 30-70 percentile range. Similarly, a range for the gamma shape parameter was determined by evaluating this parameter for the 55 acid gas data sets. The gamma shape parameter was found to vary within the bounds 2 to 4. A sample size of n=260, reflecting the average size of the acid gas data sets, was chosen for the Monte Carlo experiments.

With these parameter values 100 Monte Carlo experiments were performed. The relative root mean square errors and the average relative bias were evaluated for each combination of α_k and gamma shape parameter. These results are reported in Tables 5.5 and 5.6. In order to assess the accuracy of the values reported in Tables 5.5 and 5.6, 200 Monte Carlo experiments were performed with the random number generator initiated with a different seed, with a gamma shape parameter of 3 and setting α_k = 0.0005, 0.0015. The relative root mean square values were respectively 0.159 and 0.429 while the bias estimates were 0.133 and 0.416. These results indicate that the numbers reported in tables are accurate to about ±0.01.

The results of Tables 5.5 and 5.6 show that for the majority of data sets considered here the effect of a changing K-factor will be considerably less than a factor of 2. In fact for the fit to the 30-70 percentile range the relative root mean square error for all 55 acid gas data sets was 0.290 with the relative bias being -0.0582. Tables 5.5 and 5.6 also indicate that the value of the gamma shape parameter does not significantly affect the value of the relative root mean square errors or bias when compared with the effect of variation of α_k. Also from Tables 5.5 and 5.6 the magnitude of the results are nearly equivalent for the same absolute value of α_k. Hence the effect of positive and negative values of α_k will, overall, tend to cancel out. This is reflected in the relative bias for the 55 acid gas data sets of -0.0582.

The problem of what constitutes an acceptable level of variation in the K-factor will require careful investigation for each application of the hybrid model. Certainly the magnitude of estimates of α_k will vary with sample size. For the sample size of n=260 considered here, an acceptable range of α_k values would be ±0.0010 which would result in a variance of about ±20-30% for estimates of the 98-percentile concentration. This is well within the factor of 2 normally expected of air quality models. Also, providing α_k does not exhibit a significant bias from mean zero, predictions of the 98-percentile should exhibit only a small bias relative to the variance of the estimates.

Table 5.5: Average relative root mean square errors for estimates of the 98-percentile of the gamma distribution with the K-factor varying over the 30-70 percentile range.

α_k	Gamma shape parameter		
	2.0	3.0	4.0
0.0015	0.398	0.426	0.435
0.0010	0.268	0.287	0.291
0.0015	0.148	0.157	0.153
0.0000	0.090	0.079	0.062
-0.0005	0.167	0.156	0.149
-0.0010	0.281	0.275	0.270
-0.0015	0.398	0.389	0.375

Table 5.6: Average relative bias for estimates of the 98-percentile of the gamma distribution with the K-factor varying over the 30-70 percentile range.

α_k	Gamma shape parameter		
	2.0	3.0	4.0
0.0015	0.378	0.412	0.426
0.0010	0.244	0.269	0.279
0.0015	0.110	0.129	0.136
0.0000	0.0203	-0.0068	-0.0041
-0.0005	-0.148	-0.140	-0.139
-0.0010	-0.273	-0.269	-0.267
-0.0015	-0.393	-0.387	-0.373

Figures 5.2 and 5.3 indicate that the 20-80 percentile range may provide improved estimates of the 98-percentile as the increased number of data points would lead to a decrease in the uncertainty associated with estimates of the parameters of the gamma distribution. Accordingly, the average relative sum of squares error and average relative bias were evaluated for the estimates derived using the 20-80 percentile range. The sum of squares error was 0.344 while the bias was -0.140. These results demonstrate that the performance based upon the 30-70 percentile range is better than that for the 20-80 percentile range. This result is further supported by the correlation coefficients obtained from the fit of model predictions against observations. For the 30-70 percentile range the correlation coefficient was 0.707 while for the 20-80 percentile range this value was 0.674.

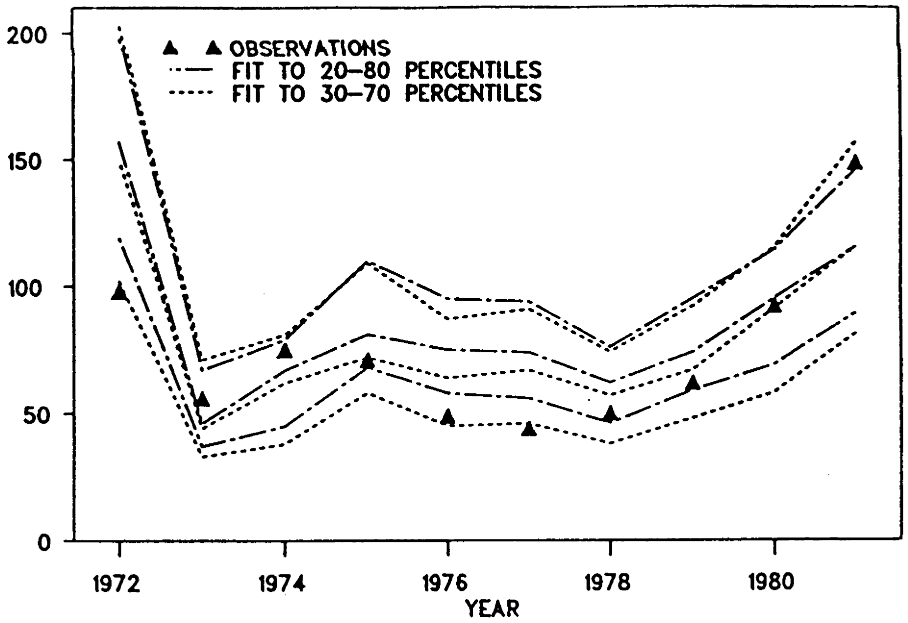

Figure 5.10: Hybrid model estimates of the 98-percentile acid gas concentration for the Watt Street data using the 30-70 percentile range (—–) and the 20-80 percentile range (---) with 95% confidence intervals indicated.

Why the 30-70 percentile range produces better model estimates of the 98-percentile than the 20-80 percentile range is considered to be due to the increasing uncertainty associated with the estimates of the 20-30 and 70-80 percentiles produced by equation (5.2). This uncertainty counters the advantage of estimating the gamma model parameters from the larger percentile range. Using the 20-80 percentile range reduces the size of the 95% confidence interval because of the increased sample size. Figure 5.10 illustrates, for the Watt Street data, model estimates and approximate confidence intervals based upon the 30-70 and 20-80 percentile ranges. Figure 5.10 shows that the 95% confidence intervals are reduced when fitting the 20-80 percentile range. However, for the confidence intervals based upon the 20-80 percentile range more observations have been excluded. Again the uncertainty associated with concentrations predicted by equation (5.2) over the more extreme percentiles has resulted in poorer performance of the hybrid model when estimating the confidence intervals.

5.6 Discussion

Simpson and Jakeman (1984) were able to demonstrate how a hybrid model could be applied to estimate the effects of fluctuations in long term meteorology on observed maximum acid gas levels. Their model was restricted to 2 of a possible 6 data sets by the assumption that both windspeed and pollutant concentration are required to be lognormally distributed.

Their results indicated that fluctuations in maximum concentrations of the order of a factor of 2-4 could be attributed to the effect of long-term meteorological change. The unexplained variation of the order of a factor of 2 was considered to be incorporated in the variation in the estimate of the K-factor, which as is given in equation (5.1), is dependent on the area source strength and atmospheric stability.

For the 98-percentile concentration a similar variation to that obtained by Simpson and Jakeman (1984) for the maximum concentrations is evident. This fluctuation may be observed for all acid gas monitoring sites within Newcastle over all years for which records are available. As might have been expected, the variation in 30-70 percentile range due to factors other than meteorological change would appear to be reduced over that for the maximum concentration. Rather than being about a factor of 2 this variation is about ±30% and is attributed to changing source strength and to a lesser extent to varying atmospheric conditions. It should be noted, however, that this variation may in fact be smaller than hypothesized as the confidence intervals derived for each model estimate are theoretical and should only account for the uncertainty in the estimation of the gamma model parameters from the sample data within $30 < p < 70$ percentiles. In section 5.5 it was shown that uncertainty in the estimate of K would have a relatively small effect on the estimate of the gamma scale parameter. Thus it would be expected that the 95% confidence levels are representative of the variation in the composite model estimate of the X_{98}. That some 85% of the estimated X_{98} values fall within the 95% confidence interval, with most of those falling outside this interval doing so by only a small margin, would indicate that the hybrid model may explain more of the variation than originally hypothesised by Simpson and Jakeman (1984).

Furthermore, a simple analysis of the scatter of observed values within the 95% confidence region reveals that no obvious systematic error is apparent. However where the assumption of a constant K-factor over the 30-70 percentile range of the gamma model is not appropriate, as noted in the previous section, then the observed 98-percentile may fall outside the 95% confidence interval. This is the case for the observations of acid gas at the Watt and Mounter sites in 1972.

5.7 Conclusions

This chapter has refined the hybrid urban model of Simpson and Jakeman (1984), making it more flexible in application and providing a methodology for quantifying the effects of uncertainty associated with model predictions. Having stressed the importance of using powerful tests of identification of distributional forms, the gamma distribution was chosen as the relevant representation for the 55 years of data available. With this distribution as the statistical component of the hybrid model, it was seen that at least 85% of the predictions of the 98-percentile concentration fall within a 95% confidence interval. The 85% figure is conservative because obvious cases where the assumptions of the model are not satisfied a priori have not been omitted from consideration.

Clearly the approach could be further refined and extended if more comprehensive data ~~were~~ was available. With more detailed information on emission strengths and meteorology, a more sophisticated deterministic component could be used which should produce a concomitant reduction in the uncertainty of hybrid model predictions, and the meaning of the K-factor could be interpreted in terms of emissions and meteorological factors such as atmospheric stability.

CHAPTER 6

A HYBRID MODEL FOR PREDICTING THE DISTRIBUTION OF POLLUTANTS DISPERSED FROM ROADWAY LINE SOURCES

6.1 Introduction

Many mathematical models have been developed to predict the dispersion of inert pollutants from roadways. These models employ a range of techniques from the simple ATDL assumption at one extreme (Hanna, 1978), often the Gaussian line source assumption (Zimmerman and Thompson, 1975), statistical models aimed at detecting trends (Tiao and Hillmer, 1978) and through to computationally complex solutions based upon the conservation of pollutant mass (Maddukuri, 1982). In this chapter the best features of a deterministic model are combined with those of a suitable statistical model to obtain estimates of pollutant concentration over the entire range of its distribution.

The deterministic model used is the General Motors (GM) model developed by Chock (1978). This model is used to predict carbon monoxide (CO) concentrations about the median. The model output is then used to estimate parameters of a suitable statistical model from which the probability of the occurrence of extreme pollution episodes can be evaluated. For the CO data, the Weibull frequency distribution was identified from a range of alternatives as an appropriate statistical model using the selection procedure developed in Chapter 4. The hybrid model developed in this study was calibrated using data recorded near a roadway line source. A model validation exercise was performed using data obtained at a second roadway site 7 km distant from the model calibration site. Good agreement was found between the hybrid model predictions and observations over the entire distribution of pollutant concentration.

6.2 Line source dispersion models

The calibration of deterministic models for line sources is often based on a limited number of valid data sets (Chock, 1982a). For example, the calibrations performed by Chock (1978) and Sistla et al. (1979) were based on the dispersion of known amounts of sulphur hexafluoride (SF_6). For the dispersion of SF_6 , these models were found to give high correlations between the predicted and observed concentrations. In particular, the GM model (Chock, 1978) was found in an independent review of several models to yield the most reliable estimates of SF_6 concentrations (Sistla et al., 1979). It was found that the Gaussian models performed at least as well as the numerical models and in the case of the GM model, often better.

Sistla et al. (1979) found that the GM model reproduced the observed vertical dispersion parameter for SF_6 observed in this experiment. In the GM model this parameter represents the dispersion of pollutant concentration with increasing distance, wind-road angle and atmospheric stability. It was concluded that SF_6 gas may be employed to validate models for CO on the basis of the similarity of concurrent measurements of the vertical dispersion of CO and SF_6 (Sistla et al., 1979). It is noted for all the data (n=108) that the correlation coefficients between predicted and observed concentrations for the various models ranged from 0.48 to 0.80. The high correlation coefficient for the GM model predictions is consistent with the level of correlation obtained in the calibration of this model by Chock (1978). From these results it may be reasonably concluded that the GM model adequately describes dispersion from roadways when all input variables are accurately measured over appropriate time scales.

In a more recent review (Rodden et al., 1982) the Gaussian plume models CALINE-3, CALINE-2, AIRPOL-4A, HIWAY and TRAPSIIM were compared on the basis of predicting the pollutant concentrations obtained from five experiments including the one used to validate the GM model. The remaining four experiments represent the observation of CO concentrations near roadways recorded as 15 minute averages. No single model yielded a clear 'best' result. For the majority of data sets, to which all models were applied, the correlation coefficients fell below 0.5. Plots of the predicted concentrations against experimental observations showed no strong correlation over the entire range of pollutant concentration. All models were found to overpredict in the range below 0.5-1.5 ppm while the higher values were underpredicted (Rodden et al., 1982). In general for the CO data only 50% of the predicted CO concentrations fell within 1 ppm of the observed concentration. Such results, while not including the GM model performance, indicate the increased uncertainty in prediction which is produced when comprehensive monitoring of windspeed, wind direction, and atmospheric stability do not take place and the source strength of the pollutant must be estimated. When normal monitoring is carried out, as distinct from intensive experimentation, these problems will usually arise. Chock (1985b) notes the very serious problems of comparing model performance using CO where there exists uncertainty in the emission rate and the background CO concentration. Such difficulties are apparent when, as was noted by Rodden et al. (1982), negative pollutant values are obtained after subtracting upwind from downwind concentrations. These results were attributed to the fluctuation of windspeed and wind angle occurring over the 15 minute averaging time. Clearly where longer averaging times are involved this problem will be amplified.

Green and Bullin (1982) also found that the three models CALINE-2, HIWAY and AIRPOL-4 were unable to reproduce the variation of mass flux profiles with height. They also demonstrated that these dispersion models were inaccurate due to the assumption of a constant windspeed with height and as a result of incorrectly representing dispersion when the wind angle to the roadway differs from 90°.

Studies of the dispersion of CO by Watson (1983), and particulate lead by Mainwaring and Thorpe (1983), applied the GM model to predict pollutant concentration. Both studies found that the GM model provided excellent agreement with observed pollutant concentrations. However these studies examined small data sets (n=24, n=44). Watson (1983) examined the performance of the GM model in predicting pollutant observations obtained at the road edge where the line source had an accompanying self-generated turbulence, a region in which the GM model may not be applicable. Benson (1982) discusses modifications to the Gaussian vertical dispersion parameter σ_z under these conditions.

Rao and Visalli (1981) compared the performance of air quality models on the basis of paired and unpaired observations. They examined the ability of four line source dispersion models, HIWAY-1, HIWAY-2, AIRPOL-4, and CALINE-3 to predict the GM model SF_6 experimental data as both paired and unpaired observations and concluded that an exponential model will describe the upper percentiles of the distribution of pollutant concentration. Rao and Visalli (1981) noted the unknown sensitivity of the upper percentiles of the distribution to variation given that the results of paired analysis indicated the models were estimating the upper percentiles for the wrong physical reasons. This point is particularly emphasised by Chock (1982b) who recognises the importance of pair-wise analysis of air pollutant data in order to establish where deterministic air quality models provide the most accurate estimates and where the weaknesses exist within these models. Clearly such an analysis should eventually lead to a model capable of accurate prediction of the upper percentiles of pollutant concentration, but such a result has yet to be forthcoming. Nevertheless, the paired comparison approach does allow an assessment of central tendencies or average situations (Rao and Visalli, 1981).

6.3 Data set for the hybrid model calibration

The data set examined consists of hourly average measurements of CO concentration recorded at a height of 3.5m near a freeway in the surburban area of Melbourne, Australia, by the Victorian Country Roads Board (Maccarrone, 1985). The section of freeway is a 1km straight segment running approximately north-west to south-east. The roadway has two lanes in each direction separated by a wide median strip and has an average daily traffic volume in excess of 30,000 vehicles. Figure 6.1 illustrates the location of the air quality monitor relative to the roadway. Meteorological information including windspeed as an hourly average and wind direction as an hourly sector average were determined at a height of 10m at the CO monitoring site. Atmospheric stability categories were recorded according to the Pasquill-Turner classification. Data sets were obtained through the months of November 1981, February 1982 and March 1982 but the data set for February 1982 consists only of four days of records.

Motor vehicle traffic counts were determined in both directions on an hourly basis for several days each month. The average traffic counts for each hour of the day and for each carriageway were determined. Table 6.1 presents the mean and standard deviation of the traffic counts over the three months. In general the standard deviation falls within 15% of the mean. These mean values were employed in the prediction of pollutant concentration. A traffic count during a working weekday yielded an estimate of the percentage of heavy duty vehicles as 6.6% of the total vehicle count.

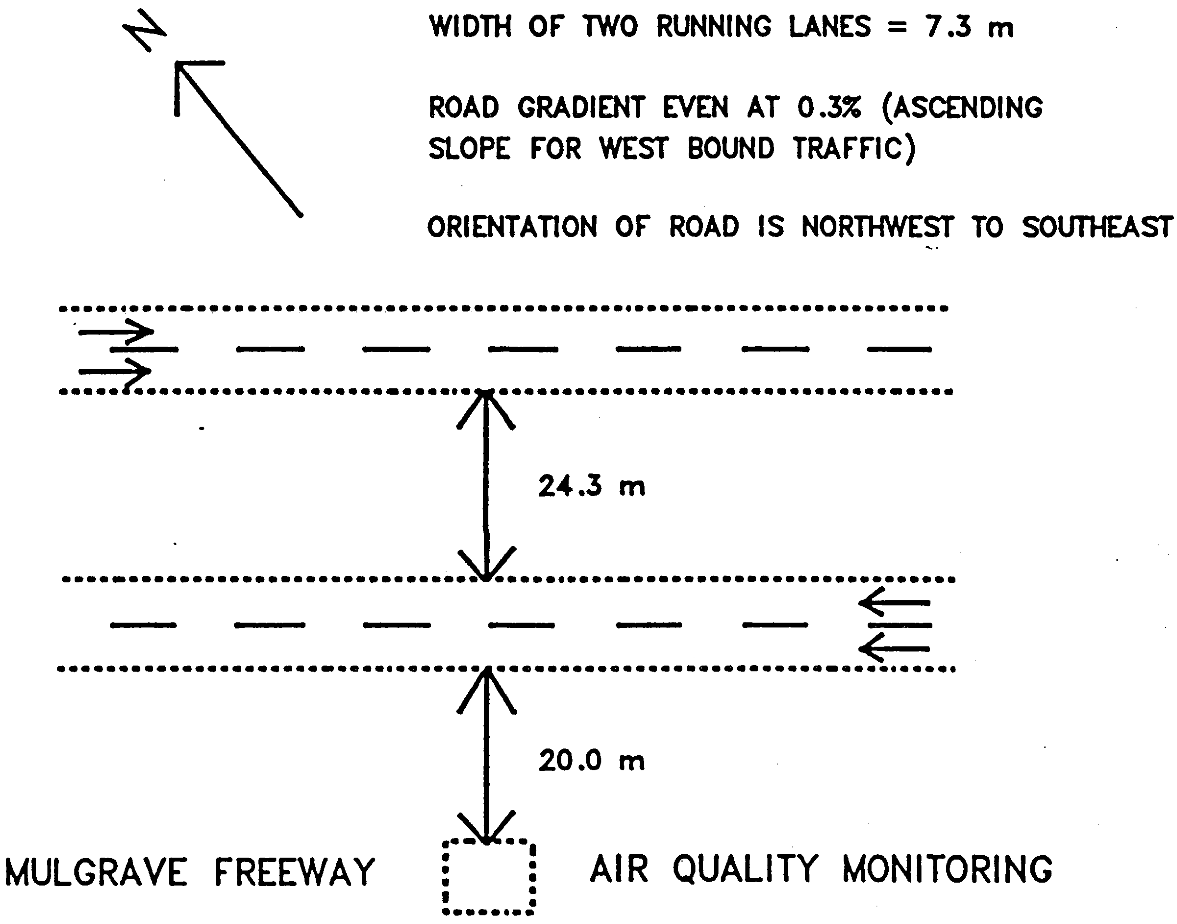
N
WIDTH OF TWO RUNNING LANES = 7.3 m
ROAD GRADIENT EVEN AT 0.3% (ASCENDING SLOPE FOR WEST BOUND TRAFFIC)
ORIENTATION OF ROAD IS NORTHWEST TO SOUTHEAST
24.3 m
20.0 m
MULGRAVE FREEWAY
AIR QUALITY MONITORING

Figure 6.1: Location of the air quality monitoring equipment in relation to the roadway for the model calibration data collection.

Table 6.1: Average traffic counts.

Hour	Nearside traffic Mean traffic count	 Standard deviation (%)	Farside traffic Mean traffic count	 Standard deviation (%)
1	140	14 (10)	245	92 (38)
2	91	17 (19)	133	60 (45)
3	80	15 (19)	81	28 (36)
4	98	23 (23)	71	18 (25)
5	133	17 (13)	89	20 (23)
6	269	32 (12)	169	20 (12)
7	1162	215 (18)	452	48 (11)
8	2460	65 (3)	1008	130 (13)
9	2266	253 (11)	1451	104 (7)
10	1214	229 (19)	973	87 (9)
11	961	101 (11)	868	88 (10)
12	897	69 (8)	845	100 (12)
13	824	78 (9)	805	59 (7)
14	836	98 (12)	792	56 (7)
15	884	76 (9)	874	41 (5)
16	1104	14 (1)	1195	63 (5)
17	1488	67 (5)	1964	151 (8)
18	1416	126 (9)	2006	58 (3)
19	1034	193 (19)	1471	230 (16)
20	782	123 (16)	942	150 (16)
21	522	72 (14)	669	143 (21)
22	378	71 (19)	527	113 (21)
23	344	48 (14)	490	15 (3)
24	257	56 (22)	394	83 (21)

In previous tests of the GM model the input parameter about which there is least certainty is the hourly average source strength of CO. Kent and Mudford (1978) have derived an expression for the CO emission rate from the performance of 28 vehicles in Australian city driving cycles and concluded that CO was emitted at a rate of $465S^{-0.97}$ g/km where S is the vehicle speed in km/hr. It should be noted that this expression was considered applicable in the range $10<S<70$ km/hr. In this thesis for the hybrid model calibration the average vehicle speed was estimated as 97km/hr. However an examination of the data of Kent and Mudford (1978) indicated that the CO emission rate was asymptotically approaching a minimum emission rate of approximately 6.0g/km with increasing vehicle speed. In a more recent study based on Australian and United States experience, Johnson (1980), gave separate emission rate estimates for CO from light duty vehicles as $662S^{-0.85}$ g/km and for heavy duty vehicles as $1220S^{-0.85}$ g/km. Given that around 6.6% of the total vehicle count comprised heavy duty motor vehicles the results of Johnson (1980) were applied in this study.

6.4 The GM model

The GM model (Chock, 1978) was developed under the assumption of an infinite line source pollutant emission regime. This assumption avoids the problem of the usual Gaussian line source model which is based on the superposition of the contributions from all the infinitesimal point sources making up the line source and thus requires numerical integration over the length of the line source considered to contribute to the pollutant concentration at some point away from the line source. The GM model may be stated as follows:

$$\chi(x,z) = \frac{Q}{\sqrt{2\pi}U\sigma_z}\{\exp[-\tfrac{1}{2}(\frac{z+h_0}{\sigma_z})^2] + \exp[-\tfrac{1}{2}(\frac{z-h_0}{\sigma_z})^2]\} \qquad (6.1)$$

where z is vertical height, $\chi(x,z)$ is the concentration at the point (x,z) relative to the line source at x=0 where the x-direction is perpendicular to the line source, Q is the emission rate per unit length, U is the windspeed perpendicular to the roadway, σ_z is the vertical dispersion parameter and h_0 is the plume centre height at distance x from the road. The form of σ_z adopted is that given by Chock (1978) incorporating the variation of σ_z with wind direction, θ, as

$$\sigma_z = (a+bF(\theta)x)^c \qquad (6.2)$$

where $F(\theta)$ is defined as

$$F(\theta) = 1+\beta\left|\frac{\theta-90}{90}\right|^{\gamma} \qquad (6.3)$$

with a, b, c, β and γ determined for neutral, stable and unstable conditions by Chock (1978).

Included in the windspeed measurement is a correction for the effective advection of pollutants due to the wake generated by the traffic (Chock, 1978). When the crossroad windspeed is $>1ms^{-1}$ the effect of plume rise is negligible and thus h_0 corresponds to the source height, otherwise the height of plume z is determined as (Chock, 1978)

$$z = (\frac{F_1}{\alpha U'^3})^{1/2} x \qquad (6.4)$$

where $F_1 = 0.052m^3S^{-3}$, x is the distance from the roadway, $U' = U_a + U_1$, U_a is the ambient crossroad wind, U_1 is a windspeed correction, and α is a constant.

Since windspeed was measured as a 45° sector average, pollutant concentrations were evaluated according to

$$\chi_S = \frac{1}{\Delta\theta} \int_{\theta}^{\theta+\Delta\theta} \chi(x,z)\, d\theta \qquad (6.5)$$

where $(\theta, \theta+\Delta\theta)$ define the bounds of the sector over which the average concentration is determined. The importance of the application of equation (6.5) is a result of the nonlinear variation of concentration with θ. As θ approaches 0° or 180° small changes in θ produce large concentration changes. Clearly were an average value of θ used in equation (6.1) near 0° or 180°, the resulting estimate of concentration might vary significantly with small changes in θ.

The windspeed data collected at the monitoring site were obtained at a height of 10m. These data have been corrected to a height of 3.5m corresponding to the CO monitor height according to (Stern, 1976)

$$\log (U/U_0) = (1/r) \log (Z/Z_0) \qquad (6.6)$$

where Z_o is the height at which the windspeed, U_o, is known. $r = 4.5$ is taken as an appropriate constant for surface roughness of towns and city outskirts (Mainwaring and Thorpe, 1983). This yields the following expression relating windspeed at 10m to that at the CO monitor height

$$U = 0.79\ U_o \qquad (6.7)$$

6.5 The Weibull distribution

The Weibull probability density function $f(\chi)$, where χ denotes pollutant concentration, is given by the equation

$$f(\chi) = (c\chi^{c-1}/b^{c})\ \exp\ [\ -(\chi/b)^{c}] \qquad (6.8)$$

where c is the shape parameter and b is the scale parameter. Using test statistics of the Kolmogorov type and the maximum of the log likelihood function as discussed in Chapter 4 the observed carbon monoxide concentrations were tested against the hypothesis of being sampled from the Weibull, exponential, gamma, lognormal and normal distributions. The results are given in Table 6.2 and indicate the applicability of the Weibull distribution to the carbon monoxide data in this study. In Table 6.2 D_n represents the maximum difference on the probability axis between the empirical distribution function and the hypothesized distribution function. The term T_n represents the value below which 95% of D_n values would fall if the hypothesized distribution is the actual distribution from which the samples were drawn. Hence the lower the value of the ratio D_n/T_n the higher the probability that this is the correct distributional model.

Table 6.2: Results of the statistical tests of the CO data set.

Model	D_n	95% Confidence level (T_n)	D_n/T_n	Log likelihood
exponential	0.168	0.060	2.80	-554.79
normal	0.129	0.049	2.63	-1231.37
lognormal	0.114	0.049	2.33	-556.68
gamma	0.0517	0.0486	1.064	-527.22
Weibull	0.0502	0.0474	1.059	-524.36

In order to predict the upper percentiles of the CO distribution a two-stage procedure was employed. Initially the GM model was used to generate estimates of pollutant concentration. These concentrations were then combined into a sample probability distribution. Hypothesising that the GM model predicts the middle percentiles with greatest accuracy, the parameters of the Weibull distribution were estimated by fitting these percentile points to the sample distribution.

The estimation procedure employs the method of maximum likelihood to evaluate the parameters of the Weibull distribution for the 30-70 percentiles. When the first r smallest values and (n-m) largest values are censored yielding the (m-r) order statistics $\chi_{r+1}, \chi_{r+2}, \ldots, \chi_m$ from a total sample of size n, the natural logarithm of the likelihood function is known as (Lemon, 1975)

$$L = (m-r)\ (\ln c - c\ln b) + (c-1) \sum_{i=r+1}^{m} \ln \chi - \sum_{i=r+1}^{m} (\chi /b)^{c}$$

$$- (n-m)\ (\chi_m/b)^{c} + r \ln[1-\exp\{-(\chi_{r+1})^{c} /b^{c}\}] \qquad (6.9)$$

where b and c are the Weibull scale and shape parameters respectively. The estimated concentrations for a given pair of Weibull parameters are derived from the inverse Weibull distribution function which is given by

$$P_i = b \{\log [1/(1-\alpha_i)]\}^{1/c} \qquad (6.10)$$

where the α_i represent the probability of some value drawn from the Weibull distribution satisfying $Pr(\chi<P_i)$. The probability values may be calculated using the the empirical distribution function which relates the order statistic to probability as

$$\alpha_i = Pr\ (\chi<P_i) = (i-0.5)/n \qquad (6.11)$$

where n is the total sample size and i represents the i-th order statistic (for example, Chambers et al., 1983).

An iterative search algorithm was applied to determine the maximum value of the likelihood function and hence the parameters of the Weibull distribution. Based upon the maximum likelihood estimates the inverse of the information matrix, the variance-covariance matrix, was evaluated for each data set. In order to determine the confidence intervals for a given percentile, a Monte Carlo procedure was utilised in conjunction with equation (6.10). 200 samples of each of the Weibull distribution parameters were selected from a correlated multivariate normal distribution (Naylor et al., 1966) with the mean values being the maximum likelihood estimates of the Weibull parameters and the standard deviations being derived from the variance-covariance matrix associated with the maximum likelihood estimates. For each pair of Weibull parameters the 98-percentile was evaluated according to equation (6.10). The 200 estimates of the 98-percentile were ordered and the approximate 95% confidence interval determined.

6.6 The hybrid model calibration

The GM model as given by equation (6.5) was applied to the carbon monoxide data set. The experimental program unfortunately did not include measurement of the background CO concentration. However estimates of the background CO concentrations were determined using the simple model

$$\chi_p = \frac{K}{u_{100-p}} \qquad (6.12)$$

where χ_p is the CO concentration and u_p the windspeed. The K-factor for each hour was determined according to the expression (see Simpson and Jakeman, 1984)

$$K = \chi_{50}\, u_{50} \qquad (6.13)$$

where χ_{50} is the median background CO concentration and u_{50} is the median windspeed recorded at the roadway monitoring site. The CO data set used for background data was recorded as hourly averages for November 1981 at the Dandenong collection site located some 5 km distant from the roadway monitoring site at least ½ km from any major traffic routes and within the same surburban area in Melbourne. This site is considered to

Table 6.3: K-factor derived using the Dandenong background CO data.

Hour ending	Median windspeed (ms^{-1})	Median [CO] (ppm)	K ($ms^{-1}ppm$)
7	2.42	0.73	1.77
8	2.43	0.80	1.94
9	3.09	0.77	2.38
10	3.40	0.59	2.01
11	4.02	0.40	1.61
12	4.67	0.46	2.15
13	4.78	0.40	1.91
14	4.66	0.266	1.24
15	3.47	0.166	0.58
16	3.99	0.166	0.66
17	3.70	0.178	0.66
18	3.63	0.178	0.65
19	3.36	0.199	0.67
20	3.07	0.266	0.82

provide reasonable estimates of the background pollutant concentration. The K-factors so determined appear as Table 6.3. The model, equation (6.12), of the background concentration is however inappropriate for windspeed $u<2ms^{-1}$ (Simpson and Jakeman, 1984) and thus is not applied in this range. Examination of the K-factors obtained indicates that the background concentrations are quite low. The majority of background concentrations should fall well below 2ppm.

Benarie (1980) notes that where convective mixing dominates pollutant dispersion advection would not change the concentration and thus in equation (6.12) windspeed should have an exponent near to zero. Using the background pollutant data recorded at the Dandenong site and windspeed data the power of u was evaluated using least squares analysis by taking the logarithms of both sides of equation (6.12). The value obtained for the power of u was -1.104 with a 95% confidence interval of -0.923 to -1.284. It would appear that for our data a power of -1 is appropriate for this study. This result may not however hold true for all data recorded in the Melbourne airshed as this result is based upon the analysis of the data appropriate to this thesis only.

For the purpose of this modelling exercise the data set considered has been restricted to that which the deterministic component may be applied without invalidating the assumptions on which it is based. Consequently, the modelling exercise has employed data recorded between the hours of 6 a.m. and 8 p.m. during the weekdays only. Such restrictions reduce the data set to 325 measurements of CO at the monitoring site for which all the necessary input variables are available for the GM model. This sample size is comparable with those examined by Rodden et al. (1982) and is well above that of Mainwaring and Thorpe (1983) and Watson (1983).

Figure 6.2 presents the plot of the observed CO concentrations against the predicted concentrations derived using the GM model where account is taken of the background concentration. The plot indicates the underestimation by the GM model of pollutant concentrations and the poor correlation between the predicted and observed concentrations. The correlation coefficient from the least squares fit to the data is 0.40. The diurnal variation of the observed mean CO concentration is presented in Figure 6.3 with the calibrated GM model estimates of these CO

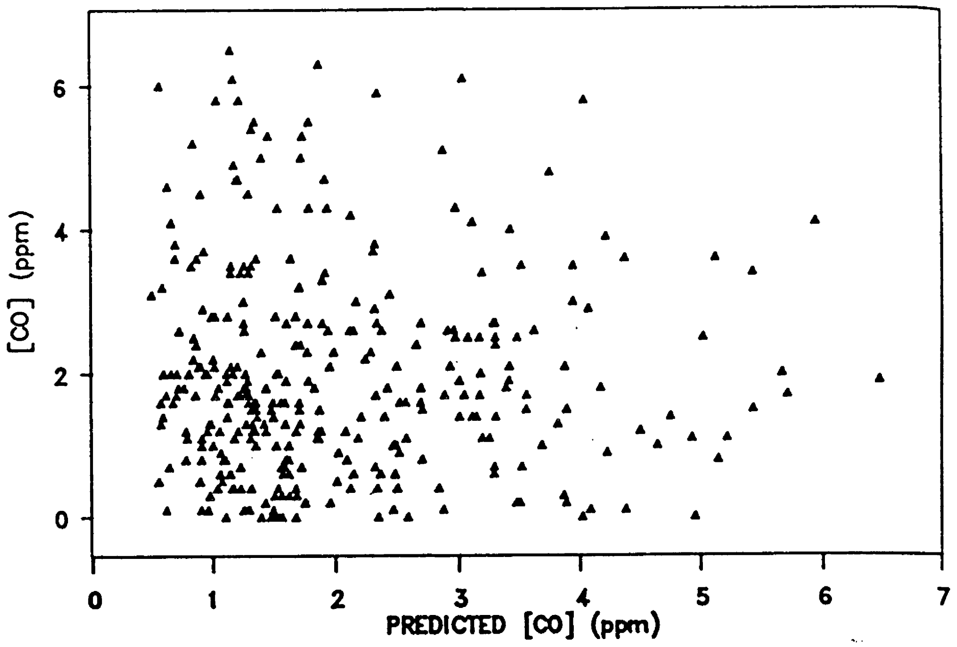

Figure 6.2: Observed versus predicted concentrations (ppm) for the model calibration data set.

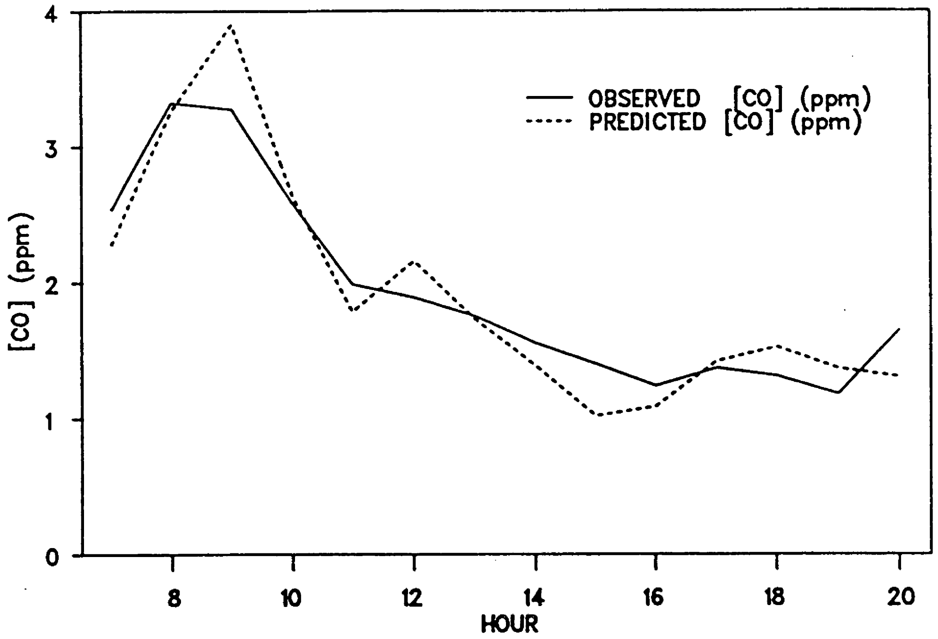

Figure 6.3: The observed and predicted diurnal variation of CO concentration (ppm) for the model calibration data set.

Table 6.4: Observed mean hourly carbon monoxide concentration (ppm) classified according to windspeed and wind direction with the GM model prediction and associated error.

	Wind angle (degrees)				
Windspeed range (ms^{-1})	0(a)	45	90	135	180
10 - 11	5.30	.00	.00	.00	.00
Model	1.01	.00	.00	.00	.00
Error(b)	18.38	.00	.00	.00	.00
9 - 10	.00	1.34	.00	.00	.00
Model	.00	.77	.00	.00	.00
Error	.00	.50	.00	.00	.00
8 - 9	.75	1.00	.00	.00	.00
Model	.78	1.10	.00	.00	.00
Error	.01	.49	.00	.00	.00
7 - 8	3.50	1.15	.00	1.50	1.50
Model	.94	1.10	.00	1.17	.62
Error	6.57	.81	.00	.12	.78
6 - 7	1.75	1.96	.00	1.40	1.60
Model	1.14	1.78	.00	.71	.77
Error	.81	.57	.00	.48	.69
5 - 6	2.80	1.55	.00	.00	1.40
Model	1.55	1.66	.00	.00	1.23
Error	2.77	.93	.00	.00	.68
4 - 5	1.10	1.44	.00	.00	1.52
Model	.87	1.99	.00	.00	1.30
Error	.05	1.24	.00	.00	1.10
3 - 4	3.00	2.44	.00	1.40	1.47
Model	2.28	3.20	.00	1.63	1.63
Error	2.34	3.42	.00	.05	1.10
2 - 3	2.87	2.92	2.50	1.77	2.13
Model	4.45	3.71	2.47	3.36	2.65
Error	4.71	3.15	2.94	3.42	2.03
1 - 2	1.90	3.36	2.73	1.75	2.83
Model	.58	2.47	1.87	2.26	1.10
Error	1.74	3.43	3.82	1.37	5.36
0 - 1	.00	1.10	.00	2.85	.00
Model	.00	2.19	.00	2.35	.00
Error	.00	1.19	.00	1.81	.00

(a) $\theta = 90$ = wind perpendicular to roadway

(b) average sum of squares error

concentrations. In this plot the GM model has been calibrated to the observed mean CO concentration to overcome the difference in scale. This calibration factor is C=3.38. Table 6.4 also presents the mean CO concentration as an average of windspeed classified in 1ms^{-1} intervals and wind direction according to the sector average. In general the mean CO value is predicted by the calibrated model to within a factor of 2 over the range of windspeed and wind direction.

Using the calibrated GM model output the parameters of the Weibull distribution were estimated from data within 30-70 percentile range using the method of maximum likelihood. The GM model output is presented on Weibull graph paper with the observed CO concentration as Figure 6.4. The close correlation of the GM model estimates to the observed concentrations over the 30-70 percentile range is apparent. It should also be noted that coincidental agreement with 70-100 percentiles has also resulted but as we shall see in the model validation section this is not always the case. The shape and scale parameters determined using the method of maximum likelihood over the 30-70 percentile range from the GM model output were b=2.118 and c=1.844 which may be compared with those estimated by the method of maximum likelihood (Thoman et al., 1969) using all the observed CO concentrations which returned values of b = 2.235 and c = 1.450. Table 6.5 gives estimates of the percentiles of the distribution of CO concentration derived using both sets of Weibull parameters. They are presented with the observed CO concentration. The resulting estimates of the percentiles of the distribution are, for the hybrid model, in excellent agreement with the observed concentrations. For the maximum concentration the approximate 95% confidence interval was evaluated as described previously. The upper and lower limits of this interval are stated with the estimated maximum in Table 6.5. The approximate confidence interval includes the observed value indicating that this approximate confidence interval provides a useful indicator of model uncertainty.

6.7 The hybrid model validation

The data for validation consists of hourly average measurements of CO concentration recorded at a height of 3.5 m near a segment of roadway some 7 km distant from the site at which the model calibration data were recorded (Maccarrone, ~~1984~~ 1985). The section of highway has two

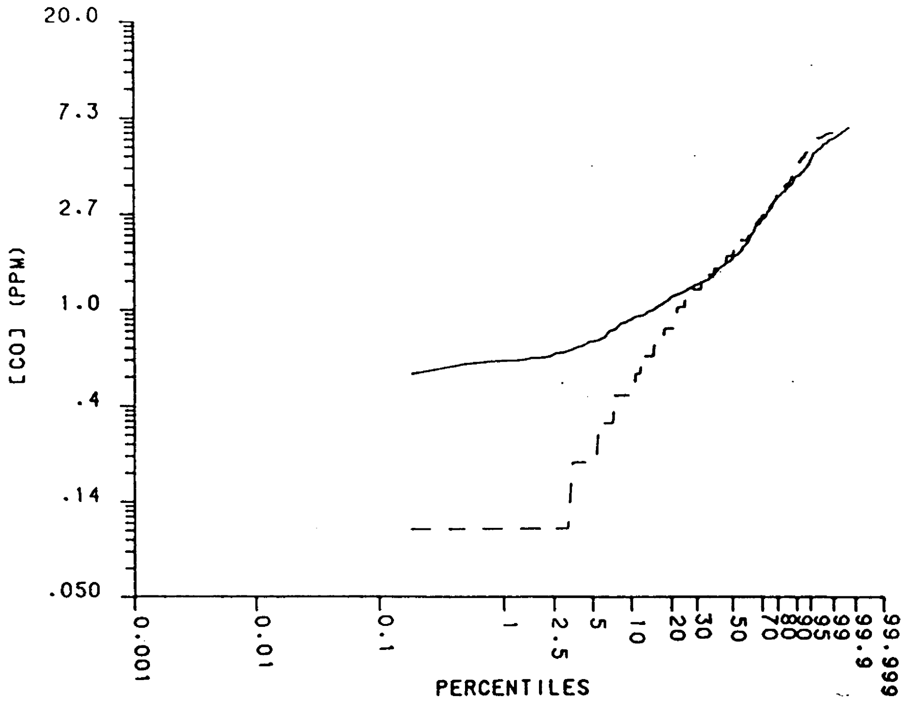

Figure 6.4: The GM model estimates (——) and observed CO concentrations (- -) as cumulative frequency distributions on Weibull graph paper.

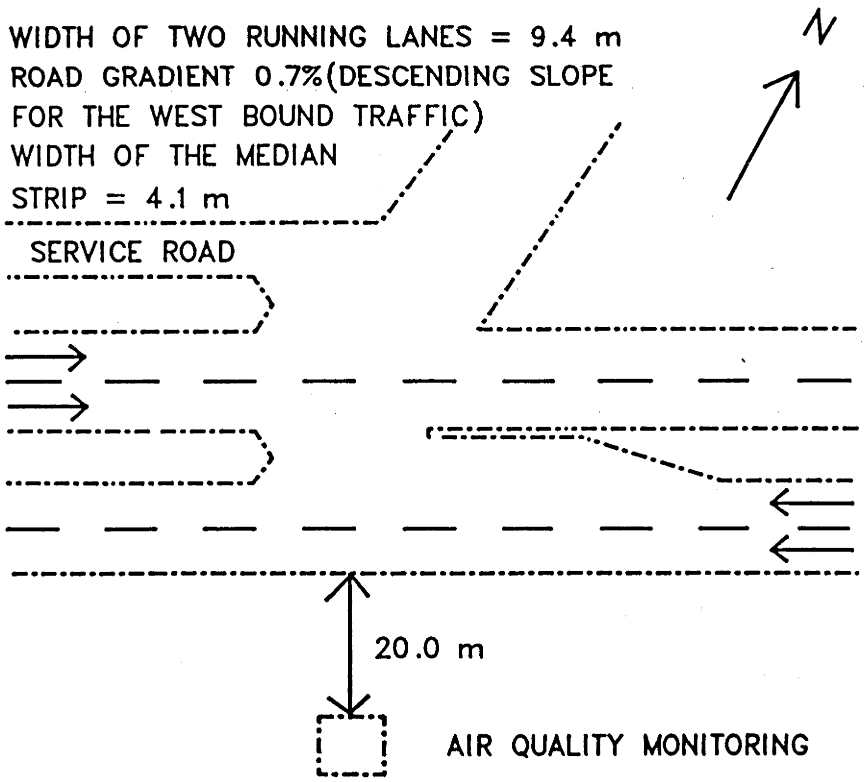
WIDTH OF TWO RUNNING LANES = 9.4 m
ROAD GRADIENT 0.7%(DESCENDING SLOPE FOR THE WEST BOUND TRAFFIC)
WIDTH OF THE MEDIAN STRIP = 4.1 m
N
SERVICE ROAD
20.0 m
AIR QUALITY MONITORING

Figure 6.5: Location of the air quality monitoring equipment in relation to the roadway for the model validation data collection.

lanes in each direction separated by a median strip with the average daily traffic volume estimated to be approximately 29,000 vehicles. The meteorological data, consisting of hourly average windspeed and wind direction as an hourly sector average, were determined at the site at a height of 10 m. Figure 6.5 illustrates the location of the air quality monitor relative to the roadway. The atmospheric stability categories, according to the Pasquill-Turner classification, were those determined on an hourly basis at the Dandenong monitoring site. The CO monitoring occurred during November 1981. However, failures of the windspeed and carbon monoxide recording instruments have restricted this data set to a smaller set of 29 observations from November 17 to 26.

Motor vehicle traffic counts were determined in both directions on an hourly basis for several days during the monitoring period. Using these data the average traffic count for each hour of the day and for each carriageway were calculated. The standard deviation of these traffic counts was on average about 15% of the vehicle count for that hour. A

Table 6.5: Hybrid model fit and maximum likelihood estimates of the observed [CO] distribution. The approximate 95% confidence interval is given for the maximum concentration.

Percentile	Hybrid model [CO] (ppm)	Maximum likelihood [CO] (ppm)	Observed [CO] (ppm)
2.50	.28	.17	.1
5.00	.42	.28	.2
10.00	.63	.47	.4
20.00	.94	.79	.8
30.00	1.21	1.09	1.2
40.00	1.47	1.40	1.5
50.00	1.74	1.73	1.7
60.00	2.01	2.10	2.0
70.00	2.34	2.53	2.5
80.00	2.74	3.10	3.0
90.00	3.33	3.97	4.1
95.00	3.84	4.76	5.0
97.50	4.30	5.49	5.8
99.00	4.85	6.40	6.1
99.84	5.08 (5.83) 7.01	7.94	6.5

traffic count during a working weekday yielded an estimate of the percentage of heavy duty vehicles as 10.7% of the total vehicle count. The average vehicle speed was estimated at 73 km/hr. As in the model calibration exercise the source strengths were evaluated using the expressions of Johnson (1980).

The GM model as given by equation (6.5) was applied to estimate the CO concentration. Background concentrations were estimated using equation (6.12) with the K-factors being those derived for the model calibration study, as listed in Table 6.3. These K-factors were derived from data collected at the Dandenong station over the same period, November 1981, as the CO data recorded at the roadway site. The calibration factor derived for the hybrid model was incorporated in the estimates of CO concentration derived here.

Figure 6.6 presents the plot of the observed CO concentration against the predicted concentrations derived using the calibrated model. The predicted mean CO concentration of 2.12 ppm is in good agreement with the observed mean CO concentration of 2.14 ppm. The model predictions

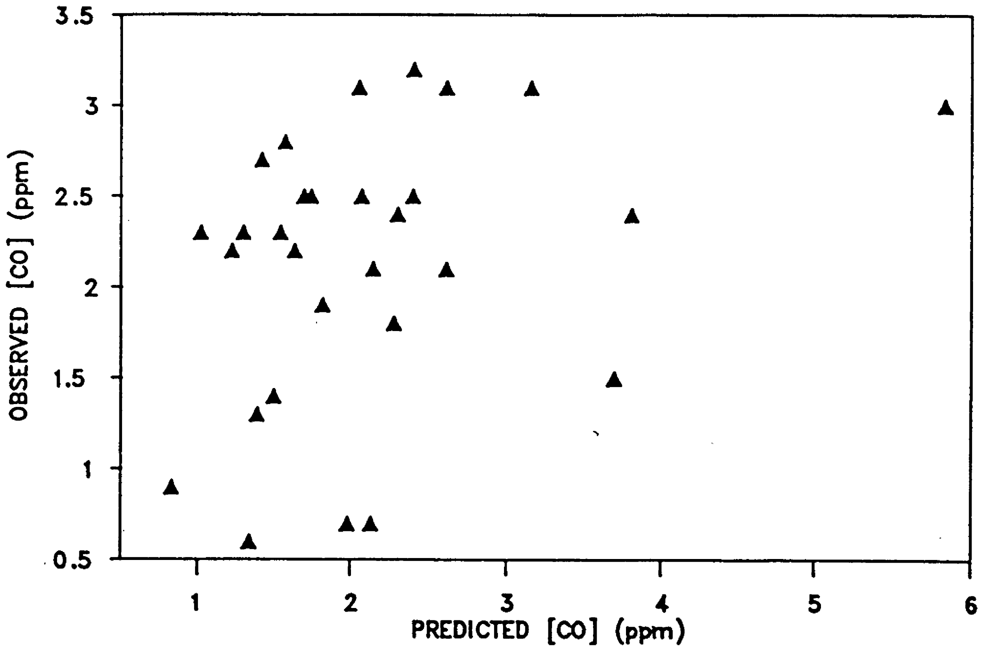

Figure 6.6: Observed versus predicted CO concentrations (ppm) using the calibrated model for the model validation data set.

Table 6.6: Observed mean hourly carbon monoxide concentration (ppm) classified according to windspeed and wind direction with the GM model prediction and associated error.

	Wind angle (degrees)			
Windspeed range (ms^{-1})	22.5[a]	67.5	112.5	157.5
6 - 7	2.57	.00	.00	2.55
Model	1.62	.00	.00	1.44
Error[b]	.94	.00	.00	1.24
5 - 6	2.25	.00	.00	2.27
Model	1.13	.00	.00	1.78
Error	1.29	.00	.00	0.41
4 - 5	2.40	.00	2.50	2.41
Model	2.30	.00	2.07	2.40
Error	.01	.00	.18	.15
3 - 4	.00	.00	.00	.65
Model	.00	.00	.00	1.74
Error	.00	.00	.00	1.30
2 - 3	1.84	.00	2.40	1.50
Model	3.91	.00	3.80	3.69
Error	4.84	.00	1.97	4.81
1 - 2	2.13	.00	.00	.00
Model	1.79	.00	.00	.00
Error	.37	.00	.00	.00
0 - 1	.00	.90	.00	.00
Model	.00	.83	.00	.00
Error	.00	.00	.00	.00

(a) $\theta = 90$ = wind perpendicular to road

(b) average sum of squares error

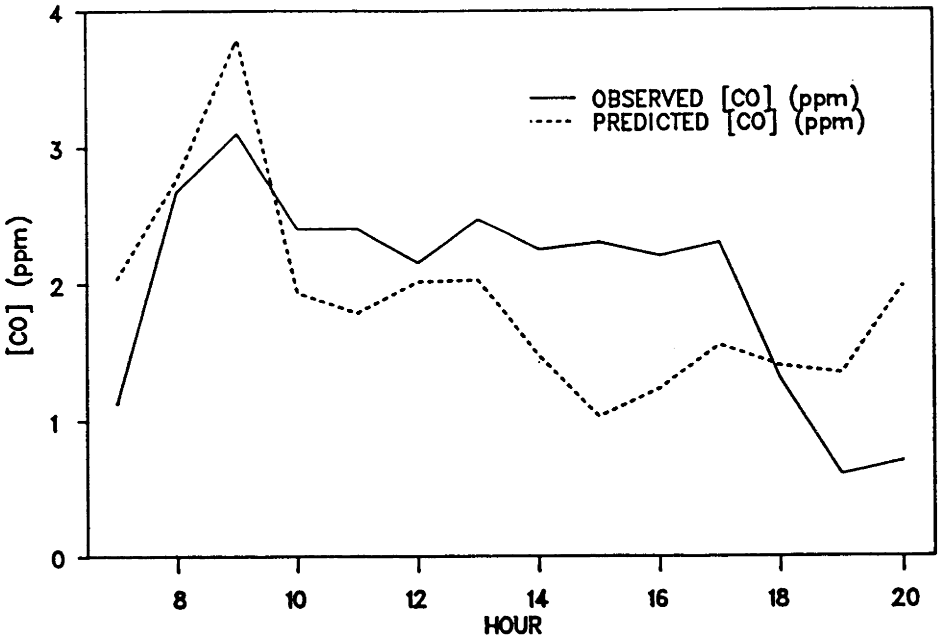

Figure 6.7: The observed and predicted diurnal variation of CO concentration (ppm) using the calibrated model for the model validation data set.

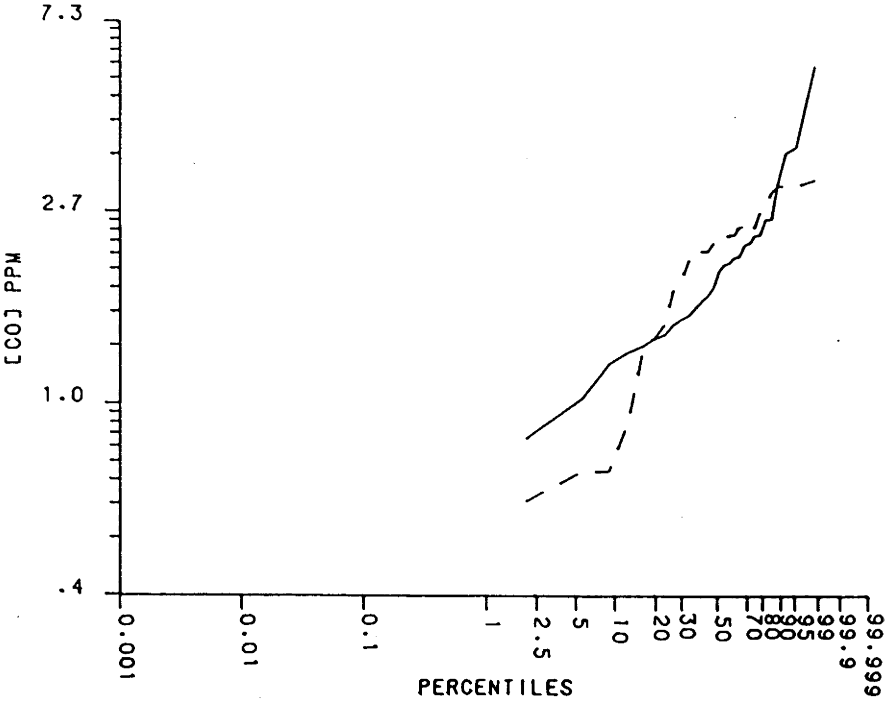

Figure 6.8: The observed CO concentrations (ppm) (- -) and GM model output (——) cumulative distributions plotted on Weibull graph paper.

were found to be within 1 ppm of the observed value 69% of the time, and within 2 ppm for 93% of the predictions. The correlation coefficient was 0.326. These results compare favourably with those obtained by Rodden et al. (1982) in their study of the performance of several pollutant dispersion models. Table 6.6 gives the mean CO concentration, predicted and observed, classified into windspeed classes of 1 ms^{-1} and wind direction according to a 45° sector average. Figure 6.7 gives the diurnal variation of the observed mean CO concentration with the corresponding model estimate. These results indicate that the deterministic component of the hybrid model predicts to within a factor of 2 the mean concentrations over a variety of conditions.

Using the GM model output the parameters of the distributional model, equation (6.8), were estimated from the data within the 30-70 percentile range using the method of maximum likelihood. Analysis of the CO validation data set using the goodness-of-fit data again preferred the Weibull model with the gamma model yielding the next best goodness-of-fit statistics. Figure 6.8 gives the GM model estimates over the entire percentile range plotted on Weibull graph paper with the observed CO concentration distribution. Agreement between the predicted and observed concentrations over the 30-70 percentile range is reasonable. It should be noted that poor correlation has resulted between the upper percentiles of the GM model output and observations. The shape and scale parameters evaluated using the method of maximum likelihood to the truncated sample over the 30-70 percentile range of the deterministic model component were $b = 2.168$ and $c = 3.303$ which may be compared with those estimated by the method of maximum likelihood from all the observed CO concentration data giving $b = 2.385$ and $c = 3.376$. Using both these sets of parameters the percentiles of the distribution of CO concentration have been calculated and are presented with the observed CO concentrations as Table 6.7. The hybrid model has produced estimates of the observed CO concentration to an accuracy of 20% over most of the range of its distribution. This may be compared with the original GM model output (in Figure 6.8) where at both extremes of the distribution overprediction of observed CO levels has occurred. For the hybrid model estimate of the maximum concentration the approximate 95% confidence interval has been evaluated and is given in Table 6.7. Again this confidence interval includes the observed maximum value indicating that the approximate confidence intervals provide useful bounds upon model uncertainty.

Table 6.7: Hybrid model fit and maximum likelihood estimates of the observed [CO] distribution of the model validation data set. The approximate 95% confidence interval for the hybrid model estimate of the maximum concentration is also given.

Percentile	Hybrid model [CO] (ppm)	Maximum likelihood [CO] (ppm)	Observed [CO] (ppm)
2.50	.71	.89	.60
5.00	.88	.99	.60
10.00	1.10	1.22	.70
20.00	1.38	1.53	1.40
30.00	1.59	1.76	1.90
40.00	1.77	1.96	2.20
50.00	1.94	2.14	2.30
60.00	2.11	2.32	2.40
70.00	2.29	2.52	2.50
80.00	2.50	2.75	2.70
90.00	2.79	3.05	3.10
98.27	2.65 (3.31) 4.65	3.61	3.20

6.8 Discussion

The GM model judged on the basis of a point by point comparison does not perform as well as in the original model calibration by Chock (1978). However this is not surprising given the uncertainty associated with model inputs, for example source strength. The results obtained are consistent with those found by Rodden et al. (1982) in their study of the performance of several Gaussian type models based on 15 minute averaged CO data. In fact the regression coefficient and the percentages of estimates within 1 and 2 ppm of the observed CO concentration are in nearly all cases above those observed by Rodden et al. (1982). Although a direct comparison using the same data sets would provide the best measure of model performance, these results are encouraging. Both Rodden et al. (1982) and Watson (1983) note the difficulties in obtaining accurate estimates of average quantities such as wind direction due to natural variations occurring over much shorter time scales than the measurement averaging time. In view of the hourly average measurements used in this study, and in comparison with the results of Rodden et al. (1982), the GM model has yielded hourly average CO concentrations to an expected accuracy of a factor of 2.

In this chapter the parameter of the GM model about which the greatest uncertainty exists is the motor vehicle emission source strength. This is illustrated by the emission factors for light duty vehicles derived by Kent and Mudford (1978) and Johnson (1980) at a vehicle speed of 70km/hr differing by more than a factor of 2. Kim and Hoskote (1983) found in a study of methods for estimating CO zonal mobile source emissions in the United States that alternative speed aggregation procedures can lead to widely divergent emission estimates. They observed that the greater the variation in link speeds within a zone and the greater the skewness of the speed frequency distribution, the greater the divergence in zonal emissions that resulted.

This uncertainty in the source strength is considered likely to have produced the high value of the calibration factor. Several factors contribute in varying degrees to this level of uncertainty. Of lesser significance is the uncertainty associated with the actual traffic volume and the percentage of heavy duty vehicles which comprise the total. For the hours 7 a.m. to 9 p.m. the variance is within about ± 15% of the average value. A similar percentage variation in the heavy duty vehicle component of the traffic volume is considered likely. In this case such variation will produce a relatively small change in source strength.

While it is desirable to obtain accurate estimates of all parameter values, the average vehicle speed is based on a limited number of observations. The average vehicle speed may vary significantly with traffic conditions and thus the high average velocity observed is likely to represent the traffic flow under optimal conditions. This is reflected in the average vehicle speed being within 3km/hr of the legal limit. Also the data of Johnson (1980) and Kent and Mudford (1978) show that increasing motor vehicle speed above 97km/hr produces only small decreases in CO emission rates. As traffic conditions may produce substantial reductions in the average vehicle speed so the emission rate of CO will increase. A drop in vehicle speed to 50km/hr would nearly double the emission rates of both heavy and light duty vehicles according to the expressions given by Johnson (1980) and Kent and Mudford (1978). They noted emission rates of 50g/km when vehicle speed is substantially below the average vehicle speed and thus under these conditions the GM model will underpredict the concentration. If average vehicle speed fell below 97km/hr on a number of occasions this might produce the calibration factor of 3.38.

It should be noted that the vehicle emission rates are based on small samples of the total vehicle population and thus the expressions derived by Kent and Mudford (1978) and Johnson (1980) may not fully represent the complex mixture of motor vehicles occurring on the roadway. For example vehicles in poor tune may produce more CO. For the Australian heavy duty motor vehicle population the precise emission levels for CO are not known. The emission factors of Johnson (1980) are based on experience in the United States and whether these emission rates are representative of the heavy duty motor vehicle population is unknown.

Given then the uncertainty in the emission factors input to the GM model it is to be expected that this model will produce best results about the mean concentration. Even though Figure 6.4 indicates a close correlation between the upper percentiles of the GM model output and the observed CO concentration distributions, the GM model is probably predicting these concentrations for the wrong physical reasons. Figure 6.2 illustrates this case from the model calibration results. For the model validation exercise, the GM model output did not provide accurate estimates of the upper percentiles of the observed CO distribution as may be seen from Figure 6.8. Thus the GM model will not necessarily provide reliable estimates of the upper percentiles of the pollutant distribution.

Instead, to predict these percentiles estimates based on the CO concentrations which the GM model predicts with greatest accuracy are derived. This region is considered to be about the mean concentration. Having identified an appropriate parametric form, in this case the Weibull distribution, the parameters of the distribution can be estimated using only the GM model data considered reliable. The accuracy of the hybrid model estimates does not depend on the fortuitous coincidence of GM model output with the upper percentiles of the pollutant distribution.

Estimates of the uncertainty associated with model estimates are derived assuming that the observations are not autocorrelated. However, as noted in Chapter 3, the CO data exhibit a significant autocorrelation. This implies that the estimates of model uncertainty developed here are conservative and represent the minimum level of model uncertainty likely to be encountered. Examination of the results for the model calibration and model validation exercises indicate that the estimated model uncertainty does provide a reasonable estimate of model uncertainty. This may be due to the incorporation of this uncertainty, at least to some extent, within the information matrix used to derive estimates of model variance. Clearly the effects of autocorrelation warrant further investigation with the view to accounting for the effects of autocorrelation upon model uncertainty.

6.9 Conclusions

In this chapter the hybrid approach has been further developed and demonstrated for describing the dispersion of carbon monoxide from a line source. Again the result has been achieved using the deterministic model output only within its range of greatest reliability, which is considered to be about the median concentration, from which the parameters of a statistical model were evaluated. Approximate confidence intervals for model predictions of the maximum concentration were also derived and were found to provide reasonable bounds upon model uncertainty in that the observed maximum concentrations fell within these bounds. While the results obtained here are encouraging further application of the hybrid model will be necessary to determine how reliable these confidence intervals are.

The deterministic component, the GM model, has yielded results comparable with other deterministic models (Rodden et al., 1982) in both the model calibration and model validation exercises. The calibration factor determined for the GM model output indicates that either more precise information concerning motor vehicle speed is required or that the estimated CO emission rates for Australian motor vehicles are higher than earlier measurements have indicated (Johnson, 1980).

The statistical model applied in the hybrid model has been identified as the Weibull model from a range of alternative distributions applicable to the study of air quality data. However it is not possible to state that this distributional model is the most useful distributional model for hourly average CO pollutant observations measured near roadway line sources in general. Also, at other averaging times, for other pollutants under different emission regimes and perhaps at new locations, statistical models besides the Weibull may be more applicable. Nevertheless, the approach here is general and allows the inclusion of other appropriately identified statistical distributions in the hybrid model.

CHAPTER 7
A HYBRID MODEL FOR PREDICTING THE DISTRIBUTION OF SULPHUR DIOXIDE CONCENTRATIONS OBSERVED NEAR ELEVATED POINT SOURCES

7.1 Introduction

Recent applications to increase emissions of sulphur dioxide arising from gold processing operations in Kalgoorlie, Western Australia have led to the introduction of a monitoring program recording sulphur dioxide concentrations in conjunction with the associated meteorological conditions. The monitoring study was to provide the necessary data for the construction of a mathematical model of pollutant dispersion. The limits on emissions were to be determined on the basis of meeting the World Health Organization goals for 24-h average sulphur dioxide concentrations. Thus a central aim was to estimate the mean and 98-percentile 24-h average sulphur dioxide concentrations. As air quality criteria have been prepared for sulphur dioxide over shorter averaging times than 24 hours (Newill, 1977) hybrid models are developed for sulphur dioxide concentrations at 8-h, 3-h, 1-h and 0.5-h averaging times also.

The hybrid models are calibrated using a full year of sulphur dioxide and meteorological data, then a model verification exercise is performed. Predicted pollutant concentrations are compared with the observations recorded at the same monitoring site and at a separate monitoring site to that of the model calibration and over different annual periods.

Jakeman and Simpson (1985) developed a hybrid model for the estimation of pollutant concentrations observed near elevated point sources. They combined the Gaussian plume model with the exponential distributional model to yield estimates of the entire distribution of pollutant concentrations. In this chapter the hybrid modelling approach for the prediction of the distribution of pollutant concentrations observed about point sources is further developed. Approximate 95% confidence intervals for percentile estimates are constructed. A point source Gaussian plume model is combined with the exponential, lognormal, Weibull and gamma distributional models to predict the distribution of 24-h, 8-h, 3-h, 1-h and 0.5-h average concentrations recorded in

Kalgoorlie, Australia. In this chapter the Gaussian plume model output is calibrated using a quantile-quantile comparison with the observations rather than as matched pairs.

7.2 The data set

The data were collected in Kalgoorlie, Western Australia. Rosher et al. (1984) describe in detail the data collection procedures for sulphur dioxide concentrations and for the meteorological parameters. A brief description of the Kalgoorlie environment and the data collection procedures follows.

Kalgoorlie lies about 360 m above sea level. The climate is hot and dry with a mean annual rainfall of about 240 mm and an annual average temperature of 26°C and minimum of 12°C. The primary sources of sulphur dioxide arise from gold roasting and nickel smelting operations using ore concentrates rich in sulphur. The four major sources of sulphur dioxide are North Kalgoorlie Mines Ltd, Gold Resources Pty Ltd, and Kalgoorlie Lake View Pty Ltd which are all located within a few kilometres of Kalgoorlie, and Kalgoorlie Nickel Smelter which is about 12 km south of the town. Two monitoring sites, one at Kalgoorlie Base Hospital and the other at Kalgoorlie Technical School, were used to obtain measurements of the sulphur dioxide concentrations. Figure 7.1 shows the location of the major sources of sulphur dioxide and the monitors within the Kalgoorlie region.

Continuous air sampling was achieved using a TRACOR 270 HA atmospheric sulphur analyser. This instrument has a cycle time of 226 seconds to analyse each ambient air sample. Any variation of sulphur dioxide concentration over a shorter time period than the cycle time will not be detected. Ten minute averages were recorded.

Meteorological parameters were recorded at both monitoring sites using instrumented 10 m towers. Windspeeds were measured using a cup anemometer with an estimated accuracy of 0.1 m sec^{-1}. Wind direction was recorded to an accuracy of three degrees. A continuous real-time measurement of the standard deviation of the wind direction (σ_θ) was also recorded with an estimated accuracy of 2.5 per cent over a measured range of 0 to 45 degrees. Unfortunately no measurements of the standard deviation of wind direction in the vertical direction (σ_e) were obtained. Air temperature measurements were recorded with an accuracy of 0.15 degrees. All meteorological parameters were averaged over a 10 minute recording period.

In this chapter the data employed were recorded from August 1, 1982 until July 31, 1983 and from January 1, 1984 to December 31, 1984 at the Kalgoorlie Regional Hospital and data were recorded at the Technical School monitoring station from March 1, 1983 until February 29, 1984 and from March 1, 1984 to December 31, 1984. The former data set was used to calibrate the hybrid model, the latter three data sets in the model validation exercise. The two data sets recorded predominantly during 1983 will be referred to as the 1983 data sets while the data sets recorded during 1984 will be referred to by that year. Table 7.1 presents the percentage of data available for analysis for the meteorological and sulphur dioxide data sets at both monitoring sites on a monthly basis. Table 7.1 shows that the available data are representative of the entire years over which data were recorded at each site. Only during the period October, November and December 1983 at the Technical School monitoring site, and in December 1984 at the Hospital monitor was data recovery particularly low. During 1984 meteorological data was recorded only at the Technical School monitor site where 100% of the data was available for all but two months when data recovery was 99.9% and 99.7% respectively. Rosher et al. (1984) noted for the Technical School monitor that the wind direction sensor performance was unsatisfactory up until 16 June, 1983 and recommended the substitution of the Hospital meteorological data set for this period. Accordingly, until the 16 June, 1983 the meteorological data employed in the modelling exercise was the Hospital data set. Rosher et al. (1984) also considered that the data recorded at the Technical School were more representative of meteorological conditions at Kalgoorlie due to the proximity of the Hospital Base Station to trees and buildings which may have affected windspeed and wind direction measurements. Unfortunately meteorological data at the Technical School is only reliably available for the period July and August 1983 for combination with the sulphur dioxide data set recorded at the Hospital monitor. As such a short time period was available, the use of the meteorological data recorded at the Hospital site was preferred.

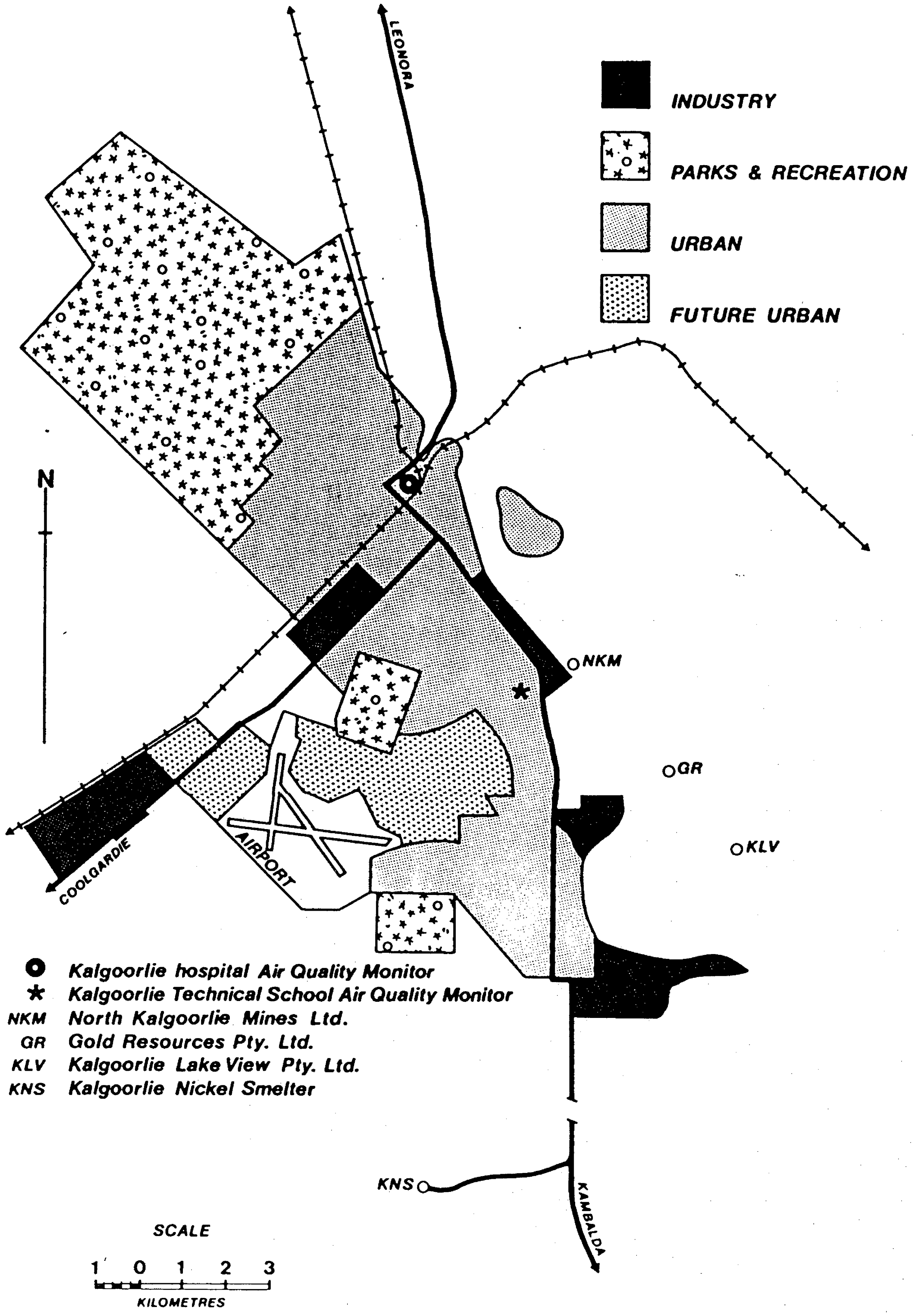
LEONORA
INDUSTRY
PARKS & RECREATION
URBAN
FUTURE URBAN
N
NKM
QR
KLV
COOLGARDIE
AIRPORT
Kalgoorlie hospital Air Quality Monitor
Kalgoorlie Technical School Air Quality Monitor
NKM North Kalgoorlie Mines Ltd.
QR Gold Resources Pty. Ltd.
KLV Kalgoorlie Lake View Pty. Ltd.
KNS Kalgoorlie Nickel Smelter
KNS
KAMBALDA
SCALE
1 0 1 2 3
KILOMETRES

Figure 7.1: Map of the Kalgoorlie region giving the locations of the major sources of sulphur dioxide and the Hospital and Technical School monitoring locations. It should be noted that the Kalgoorlie Nickel Smelter, KNS, is about 12 km from the town centre.

Table 7.1: Percentage of data available for both monitoring sites in Kalgoorlie for each month[(a)]

Month	Hospital site			Technical School site		
	Meteorological data	Sulphur dioxide data		Meteorological data	Sulphur dioxide data	
		1983	1984		1983	1984
1	100.0	98.5	36.9	51.8	50.6	-
2	100.0	97.5	71.1	100.0	99.3	-
3	100.0	98.4	99.6	100.0	99.6	84.2
4	100.0	98.7	99.3	97.5	98.3	61.7
5	100.0	98.3	98.6	100.0	99.1	99.0
6	96.9	89.3	99.4	100.0	98.2	98.8
7	96.8	90.7	93.1	99.9	92.3	98.9
8	61.1	99.1	90.1	100.0	22.2	97.5
9	100.0	97.4	22.1	100.0	66.9	57.7
10	99.8	88.3	72.1	99.8	68.4	94.3
11	100.0	98.1	99.0	99.9	90.8	20.6
12	100.0	99.1	0.0	100.0	97.3	81.4

(a) Data for 1983 adapted from Rosher et al. (1984)

Clearly there is some overlap in the meteorological data sets for the model calibration and model validation exercises for the 1983 data sets. However the model validation is also carried out at a separate monitoring site for an additional year and at the Hospital monitor over a separate yearly period to that of the model calibration. It is considered then that the model validation is adequate to test both the spatial and temporal predictive capabilities of the model.

7.3 Deterministic model component

The development of modelling methods for the prediction of the dispersion of pollutants from point sources has centred on the gradient transport and the Gaussian plume formulations (Pasquill and Smith, 1983). Hanna (1982a) in a review of deterministic models found that the predictive ability of the two approaches were similar despite differences in their plume rise and diffusion submodels. Pasquill and Smith (1983) note that the K-theory approach is restricted to situations in which the scale of turbulence is small in comparison with scale of the concentration field. Hanna (1982a) also notes that numerical instabilities arise in the solution of the gradient diffusion equation. Such numerical solutions would of course be computationally demanding. Given that the Gaussian plume formula has a simple analytic form and may be applied with meteorological data collected on a routine basis, this deterministic model form was preferred for the purpose of demonstrating the efficacy of the hybrid modelling approach for point sources.

While for this study of an elevated point source the preferred form of the deterministic model component is the Gaussian plume model, it should be noted that the hybrid modelling approach is not limited to applications only with this deterministic model. Any other deterministic model may be employed provided that the conditions under which such models are applicable are met, and that the output of these models reliably predicts some percentile range of the distribution of pollutant concentration.

The Gaussian plume model predicts the ground level concentration of pollutant χ ($\mu g\ m^{-3}$) due to a point source release at height Z_r (m) as

$$\chi\ (x,y,o,H) = \frac{Q}{\pi\sigma_y\sigma_z u} \exp\left(\frac{-y^2}{2\sigma_y^2}\right) \exp\left(\frac{-H^2}{2\sigma_z^2}\right) \qquad (7.1)$$

where x is the horizontal wind distance (m), and y is the horizontal crosswind distance (m), Q is the source strength ($\mu g\ m^{-3}$), u is the windspeed ($m\ s^{-1}$), σ_y and σ_z are the lateral and vertical dispersion parameters (m) respectively, and H is the effective height of the plume (m).

Given equation (7.1) above, two submodels must be evaluated. Values for σ_y and σ_z are required, and a model for plume rise, ΔH, where

$$H = Z_r + \Delta H \tag{7.2}$$

The submodel used in this study is based on a modification of the Briggs' (1975) formula obtained by Carras and Williams (1981) from their investigation of the Mt Isa, Australia, smelter plumes where conditions for dispersion are very similar to those at Kalgoorlie, and is given by

$$\Delta H = 1.3\ F_o^{1/3}\ x^{2/3}\ u^{-1} \tag{7.3}$$

where F_o is the buoyancy flux parameter. Carras and Williams (1981) replaced the factor 1.6, as used by Briggs (1975), by the factor 1.3. The buoyancy flux parameter is given by

$$F_o = \frac{gV}{\pi} \left(\frac{T_o - T_a}{T_o}\right) \tag{7.4}$$

where g is gravitational acceleration ($m\ s^{-2}$), V is the efflux volume in ($m^3\ s^{-1}$), T_o is the temperature of the exit gas (^oK) and T_a is the ambient air temperature (^oK).

As measurements of σ_θ, the standard deviation of wind fluctuations in the horizontal direction, are available, the approach of Hanna (1982a) was adopted which estimates σ_y according to

$$\sigma_y = \sigma_\theta \times S_y \tag{7.5}$$

where x is the horizontal wind distance (m), and S_y is defined according to Irwin (1979) as

$$S_y = (1 + 0.031\, x^{0.46})^{-1} \qquad x < 10^4 \text{ m}$$
$$S_y = 33\, x^{-0.5} \qquad x > 10^4 \text{ m} \qquad (7.6)$$

Unfortunately no data for σ_e, the standard deviation of the vertical wind direction fluctuations, were collected and thus a similar expression to that of equation (7.5) above could not be used to estimate σ_z. Instead estimates of σ_z are based upon the formula derived by Carras and Williams (1981) to describe the dispersion of the Mt Isa smelter plumes. The equation giving σ_z for daytime neutral-unstable conditions is

$$\sigma_z = 3.5\, t^{0.67} \qquad (7.7)$$

where t is the time of travel in seconds from source to receptor. An average inversion height of 1000m has been adopted as no detailed inversion height data are currently available. Following Hanna, (1982a) σ_z has been allowed to grow to no more than 0.8 of the inversion height. It should be noted that Carras and Williams (1983) and Chambers et al. (1982) also found from experimental studies that equation (7.7) provided a reasonable description for this dispersion parameter for the Liddell power station in the Hunter Valley, Australia. Equation (7.7) is, however, only applicable during the day-time hours. For night-time conditions the σ_z formulae recommended by Briggs (1973) have been applied.,

For each of the four major point sources of sulphur dioxide within Kalgoorlie the source emission characteristics necessary for application of the Gaussian plume model are listed in Table 7.2. The source to monitor bearings and distances are given in Table 7.3.

Table 7.2: Source emission parameters.

Source	Stack height (m)	Source strength (tonnes day^{-1})	Exit temperature (Kelvin)	Exit flux (m^3 sec^{-1})
1	76	50	533	20.58
2	76	30	533	20.58
3	64	100	535	10.56
4	153	840	573	95.78

Table 7.3 shows that the largest point source is about 14706m from the Kalgoorlie Hospital monitor. It is usually considered that dispersion over distances greater than 10000m is not well described by the Gaussian plume model. On the other hand this source is reasonably well separated according to the prevailing wind direction and, as will be shown, dispersion from this source has a sufficiently large impact that it may not be ignored. Fortunately the topography between the source and receptor is level, with no major obstacles such as buildings or lakes between source and receptor. Also, as will be noted in section 7.7, the

Table 7.3: Monitor to source bearings and distances.

Source	Monitor location			
	Kalgoorlie Hospital		Technical School	
	Bearing	Distance (m)	Bearing	Distance (m)
1	137.3°	2685	70.3°	653
2	137.9°	4298	120.7°	1948
3	137.6°	5535	126.9°	3157
4	174.0°	14706	179.0°	12444

estimates produced by the Gaussian plume model for this source agree with the observed concentrations to within a factor of 2. These results indicate that the Gaussian plume still provides a reasonable model for dispersion from the most distant source to the monitors in Kalgoorlie.

7.4 Statistical model component

The best distributional model with which sulphur dioxide concentrations may be described remains the subject of debate. Larsen (1971) assumed that all pollutants at all averaging times could be described using a model based upon the lognormal distribution. However Pollack (1975) considered that an exponential distribution should provide a better distributional model for pollutant concentrations observed about isolated point sources. For 24-h average sulphur dioxide concentrations recorded in the Gent region of Belgium, Berger et al. (1982) found that the gamma distribution provided a good description of the entire percentile range, while a model based upon the exponential distribution gave a better fit to the distribution of the largest pollutant observations.

In this section we examine the goodness-of-fit of the two-parameter lognormal, Weibull and gamma distributions and the one-parameter exponential to the distribution of non-zero 24-h, 8-h, 3-h, 1-h and 0.5-h average pollutant concentrations recorded at the two monitoring sites in Kalgoorlie. The method of model identification from amongst this set of alternatives was described in detail in Chapter 4. A brief description follows of the most important aspects of the model identification procedure.

For each distributional model the parameters are estimated using the method of maximum likelihood. Using these parameter estimates test statistics of the Kolmogorov type, the maximum difference between the hypothesized cumulative distribution function and the empirical distribution function, are computed. For each distributional model the ratio of the maximum observed difference to the respective 95% confidence level is evaluated. Where this ratio is < 1, the distributional model is accepted at the 95% confidence level.

In addition to the Kolmogorov statistic, the maximum value of the log likelihood functions were calculated. The model with the largest value of this function is selected as the model providing the best fit to the observations. In Chapter 5 it was found that the likelihood function selection criterion never selected the exponential distribution over the gamma or Weibull distributions, even when the underlying distribution is exponential. Where the observations follow an exponential distribution then using the likelihood function criterion should result in exponential, Weibull and gamma models being selected equally. It was considered that the preference for the gamma and Weibull distributions is due to the positive bias of the maximum likelihood parameter estimates. On the basis of Monte Carlo experiments it was recommended that where the Kolmogorov type statistic prefers the exponential model over the Weibull or gamma alternatives this model should be selected as the best distributional model.

For the 24-h, 8-h, 3-h, 1-h and 0.5-h average sulphur dioxide data sets constructed from the 10-minute average concentrations recorded at the Hospital and Technical School monitoring sites, the ratio of the Kolmogorov test statistic to the respective 95% confidence level and the maximum values of log likelihood functions are presented as Table 7.4 for each distributional model. The results indicate that neither test statistic prefers the exponential or lognormal distributional models at all averaging times examined here. For the exponential model the Kolmogorov ratio is greater than that for the gamma or Weibull models and thus these distributions should provide the better models for the entire distribution of sulphur dioxide observations.

For the 1983 Hospital 24-h average data set the likelihood statistic prefers the gamma distribution, while for the 1983 Technical School 24-h average data set the Weibull distribution is selected. Table 7.4 also demonstrates that the Kolmogorov statistic prefers the opposite models to that of the likelihood statistic for 24-h average data sets. That this should occur is attributed to the sensitivity of the Kolmogorov ratio to the effects of rounding of the sulphur dioxide measurements to within $\pm$ 2.86 μg m^{-3}.

Table 7.4: Test statistics for 24-h, 8-h, 3-h, 1-h and 0.5-h average sulphur dioxide observations recorded in Kalgoorlie, Australia.

Distributional model	Hospital Monitor (1983 data set)		Technical School Monitor (1983 data set)	
	Kolmogorov statistic	Log likelihood	Kolmogorov statistic	Log likelihood
		24-h average data		
exponential	1.4697	-748.10	2.8747	-940.53
lognormal	1.6312	-756.12	1.6327	-921.97
gamma	1.2486	-742.56	1.1124	-921.46
Weibull	0.9871	-743.16	1.5494	-919.17
		8-h average data		
exponential	1.2519	-2077.69	1.4523	-1480.39
lognormal	2.3055	-2125.24	1.7625	-1486.26
gamma	1.895	-2074.42	0.6799	-1469.62
Weibull	0.973	-2075.82	0.5579	-1469.02
		3-h average data		
exponential	3.7219	-2876.92	3.3130	-3147.56
lognormal	2.0203	-2817.95	2.1509	-3131.67
gamma	1.1673	-2802.66	1.4576	-3096.56
Weibull	0.7274	-2797.16	1.0957	-3097.05
		1-h average data		
exponential	6.7955	-6461.97	5.6621	-7338.11
lognormal	2.4217	-6219.17	3.7224	-7279.15
gamma	2.6137	-6195.85	2.6415	-7181.99
Weibull	0.8973	-6173.29	2.0131	-7187.28
		0.5-h average data		
exponential	7.4110	-10234.79	10.2704	-18353.25
lognormal	2.0436	-9966.26	4.0010	-17861.55
gamma	3.3875	-9956.95	3.0955	-17758.79
Weibull	1.4363	-9924.78	1.8765	-17728.94

Table 7.4 (cont.)

Distributional model	Hospital Monitor (1984 data set)		Technical School Monitor (1984 data set)	
	Kolmogorov statistic	Log likelihood	Kolmogorov statistic	Log likelihood
24-h average data				
exponential	1.0971	-629.90	1.4936	-645.40
lognormal	1.1440	-634.37	1.9472	-672.62
gamma	0.6638	-625.24	0.7030	-644.07
Weibull	0.6137	-624.88	0.8179	-644.87
8-h average data				
exponential	1.0101	-1178.92	1.2685	-1327.63
lognormal	1.6482	-1192.37	1.1437	-1338.43
gamma	0.7024	-1175.18	0.5972	-1323.97
Weibull	0.6416	-1175.06	0.6814	-1323.57
3-h average data				
exponential	2.1769	-2380.32	1.4692	-2521.71
lognormal	1.8036	-2389.70	2.3118	-2544.19
gamma	0.8393	-2355.31	0.7770	-2512.48
Weibull	0.5588	-2353.83	0.7447	-2512.92
1-h average data				
exponential	4.2840	-5554.23	2.7177	-5610.23
lognormal	2.1976	-5494.37	2.6727	-5644.65
gamma	1.5144	-5455.66	0.7404	-5572.38
Weibull	0.8135	-5447.11	1.0537	-5572.76
0.5-h average data				
exponential	5.8897	-9551.29	3.9051	-9198.67
lognormal	2.5964	-9391.24	3.1138	-9285.01
gamma	2.5555	-9365.68	1.0898	-9131.76
Weibull	1.0574	-9343.68	1.1144	-9131.59

The rounding produces a large step between individual measurements. This results in greater differences between the hypothesized distribution function and the empirical distribution function than would be expected were the data measured to the same number of significant figures as the computer upon which the table of quantiles of the test statistic were generated. With increasing sample size n the 95% confidence level reduces in size approximately according to $n^{-0.5}$. Hence the effect of rounding of the measured sulphur dioxide level upon the Kolmogorov ratio will be greatest at larger sample sizes where the rounding size may be much larger than the 95% confidence level. As the data of Table 7.4 show, the effect of rounding has been reduced somewhat due to the averaging of 144 ten minute averages of sulphur dioxide to produce one 24-h average.

For large sample sizes, for instance data sets consisting of half hourly or hourly averages recorded over one year, the combined effects of the large sample size decreasing the 95% confidence level and little averaging out of the rounding of the measurements will produce large values of the Kolmogorov ratio. The magnitude of the Kolmogorov ratio under these circumstances does not necessarily imply that the distributional model tested is not a good representation of observations, or that this statistic may not be used to select the best distributional model from amongst several alternative models.

The variance and covariance estimates of the maximum likelihood estimates for the parameters of the lognormal, gamma and Weibull distributions were obtained in the usual way - that is by inverting the information matrix with elements which are the negative expected values of the second order derivatives of the logarithm of the likelihood function. Accordingly for a location parameter a and scale parameter b the approximate variance-covariance matrix is as follows

$$\begin{bmatrix} -\dfrac{\partial^2 \ell n\, L}{\partial a^2} & -\dfrac{\partial^2 \ell n\, L}{\partial a \partial b} \\ -\dfrac{\partial^2 \ell n\, L}{\partial a \partial b} & -\dfrac{\partial^2 \ell n\, b}{\partial b^2} \end{bmatrix}^{-1} \approx \begin{bmatrix} V(\hat{a}) & Cov(\hat{a},\hat{b}) \\ Cov(\hat{a},\hat{b}) & V(\hat{b}) \end{bmatrix}$$

where L denotes the respective likelihood functions. As noted by Cohen (1965), this approach is valid in a strict sense only for large samples, but it may be relied upon to yield reasonable approximations to estimate variances and covariances for moderate size samples (n > 30) where the bias is small.

7.5 Hybrid model calibration

A common procedure for model calibration (Simpson and Hanna, 1981) involves the ordinary least squares fitting of model predictions against observations. This may be stated as

$$\chi_{obs} = a\ \chi_{model} + b \qquad (7.8)$$

where a and b are the usual ordinary least squares parameters, χ_{obs} is the observed concentration and χ_{model} the equivalent model prediction. It is usual to regard a as the calibration factor relating the observed and model concentrations and b as a zero correction factor. The zero correction factor would usually be expected to have a relatively smaller effect upon the model estimates of higher concentration than those at lower concentrations. A perfect fit between model estimates and observations would result in parameter estimates of a = 1 and b = 0. This method of model calibration allows prediction of when the maximum concentration will occur. As noted earlier, however, the ability of the Gaussian plume model to make such a prediction is limited. Table 7.5 presents the estimates of the parameters a and b of equation (7.8) and the correlation coefficient for the 24-h, 8-h, 3-h, 1-h and 0.5-h average Hospital monitor data sets. The correlation coefficients listed in Table 7.5 support the assumption that the question as to when the maximum concentration will occur will not be reliably answered. Ordering the calibrated model data also allows the maximum, or any percentile, to be determined. However, as was discussed earlier, the upper percentiles of the Gaussian plume model estimates do not usually correlate well with the observations. This model produces best estimates about the mean concentration.

An alternative approach to model calibration is based upon the empirical quantile-quantile plot (Wilk and Gnanadesikan, 1968; Chambers et al., 1983). As long as the number of model predictions is the same as the number of observations the empirical quantile-quantile plot is simply a plot of one sorted data set against the other. As with the unordered data set an ordinary least squares model can be employed to relate model predictions to the observations. This would mean that the two sets of sorted data are related approximately according to

$$X_{obs} = \alpha \, X_{model} + \beta \qquad (7.9)$$

If the ordered model predictions agreed exactly with the ordered observations, parameter estimates of $\alpha = 1$ and $\beta = 0$ would result. This method of model calibration does not directly relate model predictions to the observations as timewise pairs. Instead we examine the relationship between the distributions of model predictions and observations. Importantly, model calibration according to equation (7.9) still allows the maximum concentration, or any percentile of interest, to be estimated.

Figure 7.2 illustrates an empirical quantile-quantile plot of the predicted and observed 24-h average data sets for the Hospital monitor. A least squares fit according to equation (7.9) is also presented in Figure 7.2. The mean concentration is indicated and the 50-90 percentile range has been bracketed. Figure 7.2 illustrates the approximate linearity between the observed and predicted quantiles. However the upper percentiles of this plot deviates from the linear form. This is to be expected as the Gaussian plume model provides the best estimates about the mean concentration. Thus rather than calibrate using all the predicted and observed quantiles, model calibration is performed using only a limited percentile range about the concentration which the Gaussian plume model estimates best. This concentration is considered to be the mean value. For positively skewed distributions the mean concentration should fall within the 50-90 percentile range. This percentile range was preferred for model calibration. As indicated in Figure 7.2 this percentile range, denoted by the bracketed portion of the graph, is sufficiently linear that the assumption of a calibration curve given by equation (7.9) will not be seriously violated.

Table 7.5: Estimates of the parameters a and b of equation (8) and the correlation coefficient for the 24-, 8-, 3-, 1- and 0.5-h average Hospital monitor data sets, Kalgoorlie, Western Australia.

Averaging time (h)	a ± Standard error	b ± Standard error	R^2
24	0.338 ± 0.024	4.0129 ± 2.789	0.624
8	0.371 ± 0.017	2.206 ± 2.355	0.579
3	0.359 ± 0.013	3.542 ± 2.047	0.493
1	0.297 ± 0.009	7.971 ± 1.66	0.356
0.5	0.246 ± 0.006	11.522 ± 1.32	0.292

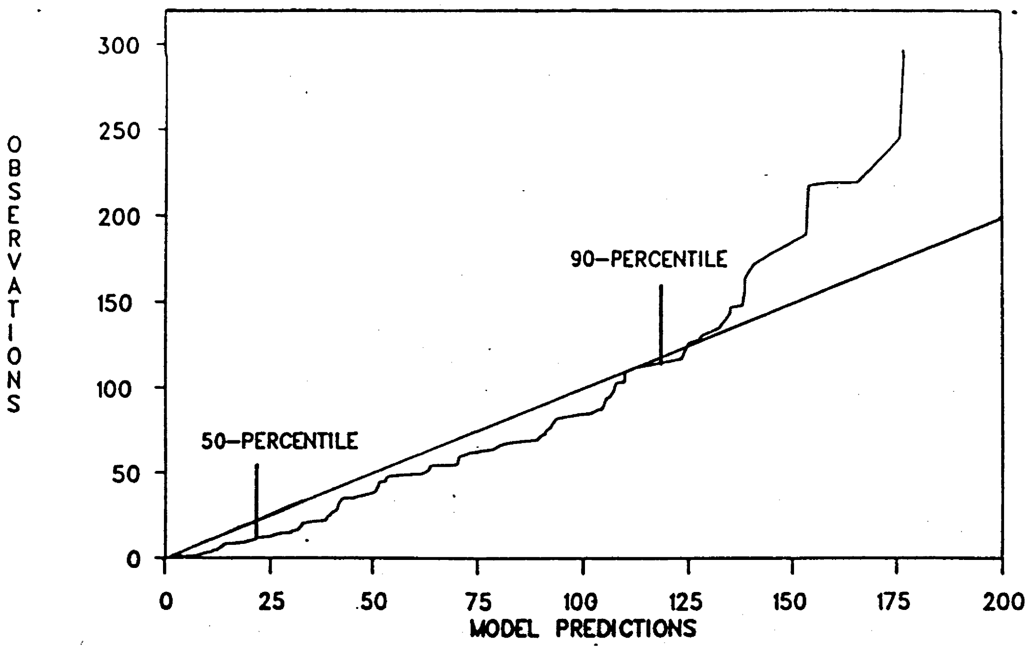

Figure 7.2: A quantile-quantile plot of the observed against predicted 24-h average data for the Hospital monitor with the 50-90 percentile range indicated.

For the Hospital monitor 24-h, 8-h, 3-h, 1-h and 0.5-h average data sets Table 7.6 presents estimates of the parameters from the model calibration according to equation (7.9) using the 50-90 percentiles. The correlation coefficients are also given in Table 7.6. These correlation coefficients indicate a very high degree of correlation between the Gaussian plume model and observations over the 50-90 percentile range. That the correlation coefficients for the 50-90 percentile range are so close to unity when compared with the results based upon matched pairs, as listed in Table 7.5, is to be expected when comparing two ordered data sets. However as all correlation coefficients are greater than 0.98 this seems to indicate that this correlation is more than just the result of the ordering of the data. It would of course be preferable to apply some goodness-of-fit test to assess whether the two distributions are the same. A test based upon the Kolmogorov-Smirnov two sample test, which would have to be modified for application to a truncated percentile range is currently under investigation.

Table 7.6: The observed and model estimate quantile-quantile calibration parameters estimated using the 50-90 percentile range for the Hospital monitor data set, Kalgoorlie, Western Australia.

Averaging time (h)	α ± Standard error	β ± Standard error	Correlation coefficient
24	0.471 ± 0.013	6.51 ± 1.84	0.980
8	0.794 ± 0.009	- 32.79 ± 1.94	0.992
3	1.059 ± 0.009	- 89.12 ± 2.29	0.993
1	1.230 ± 0.008	-154.76 ± 2.57	0.991
0.5	1.383 ± 0.009	-214.46 ± 3.39	0.986

The estimates of the parameters α and β, as given in Table 7.6, indicate that the lower percentiles of the 50-90 percentile range are over-estimated. Such a systematic correction is not apparent from the matched pair calibration according to equation (7.8) as the calibration parameters based upon equation (7.8) remain nearly constant. Using the estimates of the parameters of equation (7.9), as stated in Table 7.6, the

Gaussian plume model output was calibrated within the 50-90 percentile range. With these calibrated data the parameters of a previously identified distributional model may be estimated.

An additional advantage of model calibration according to equation (7.9), apart from calibrating a percentile range, is the improved prediction of the number of non-zero concentrations. The correct prediction of the number of non-zero concentrations is important when determining the upper percentiles of the distribution. Too many predicted concentrations will lead to overprediction, too few and underprediction will result. Thus when applying the hybrid model the number of non-zero data points must also be calibrated. In this chapter the Gaussian plume model produces more non-zero estimates of pollutant concentration than observed. Jakeman and Simpson (1985) also noted such an effect for the model employed in their study. Table 7.7 presents the ratio of the number of predicted by the Gaussian plume model, n_p, to the number actually observed, n_o , for both monitoring sites for 1983. From Table 7.7 it may readily be seen that the overprediction decreases with increasing averaging time. Using the calibration factors of Table 7.6 the number of non-zero concentrations, n_c, was reduced substantially. Table 7.7 also lists the ratio n_c/n_o which clearly demonstrates the improvement in the prediction of the number of non-zero pollutant concentrations after calibration according to equation (7.9). It should be noted that the model calibration factors determined using the Hospital monitor data were applied directly to the Technical School monitor data. This produced a similar improvement to that observed for the Hospital data set, in the predicted number of non-zero data.

7.6 Hybrid model results

Using the meteorological data sets recorded at the Hospital and Technical School monitors, estimates of the 10 minute average sulphur dioxide concentrations were calculated using equation (7.1). With these estimates 24-h, 8-h, 3-h, 1-h and 0.5-h average concentrations were constructed where it was assumed that at least 80% of the 10 minute data were available before an average over a longer time period was evaluated. From Table 7.1 it may be seen that very few averages were discarded on this basis.

Table 7.7: Ratio of the number of non-zero pollutant concentrations predicted by the Gaussian plume model (n_p), and number predicted after model calibration (n_c), to that observed (n_o) for both monitoring sites, Kalgoorlie, Western Australia.

Averaging Time	Hospital Monitor 1983 data		Technical Monitor 1983 data	
(h)	(n_p/n_o)	(n_c/n_o)	(n_p/n_o)	(n_c/n_o)
24	1.038	1.038	1.355	1.355
8	1.709	1.139	1.597	1.302
3	1.913	1.130	1.552	1.149
1	2.129	1.200	1.743	1.189
0.5	2.245	1.225	2.263	1.294

For the Hospital monitor 1983 data set the Gaussian plume model predicted a mean concentration of 69.17 μgm^{-3}. The observed mean concentration was 28.58 μgm^{-3} giving a calibration factor based on these concentrations of 0.413. This calibration factor falls within the range to be expected when applying deterministic models (Nieuwstadt, 1980; Hanna, 1982a; Benarie, 1982; Pasquill and Smith, 1983).

Figure 7.3 presents the variation of the average predicted and observed concentrations with wind direction for the Hospital monitor data. The predicted concentrations have been calibrated to the observed mean to eliminate differences in scale. The separation of the observed sulphur dioxide concentration into two distinct peaks can be seen. As may be noted from Table 7.3 this can be attributed to the three sources closest to the Hospital monitor sharing almost the same bearing. Figure 7.3 indicates that the peak concentrations fall a few degrees away from the expected wind direction. This effect may be attributed to the presence of buildings in close proximity to the Hospital monitor.

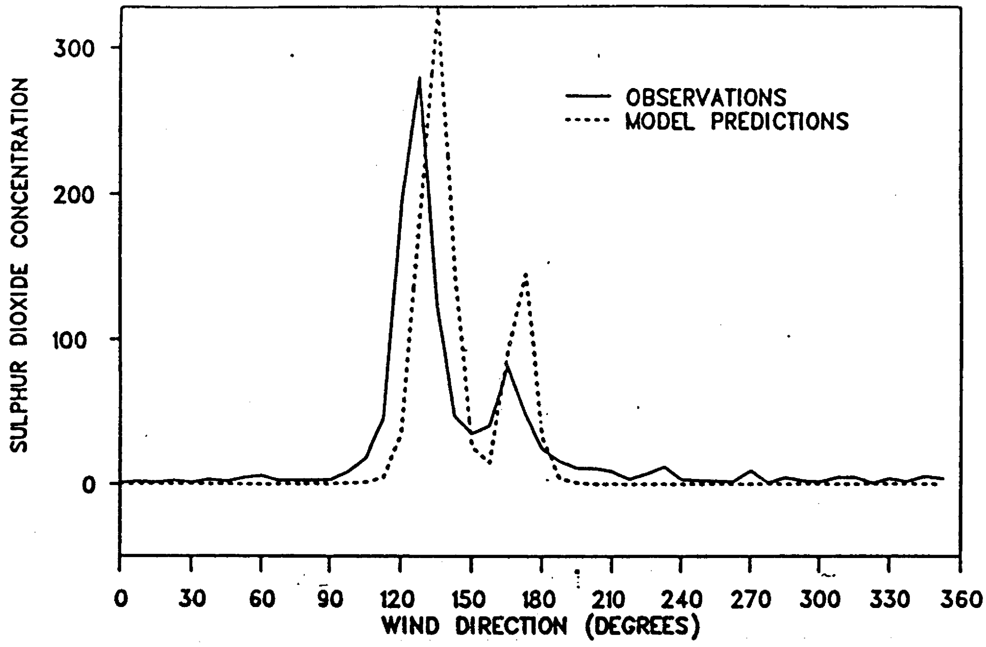

Figure 7.3: The variation of the predicted and observed average concentrations (μgm^{-3}) with wind direction for the 1983 Hospital data set.

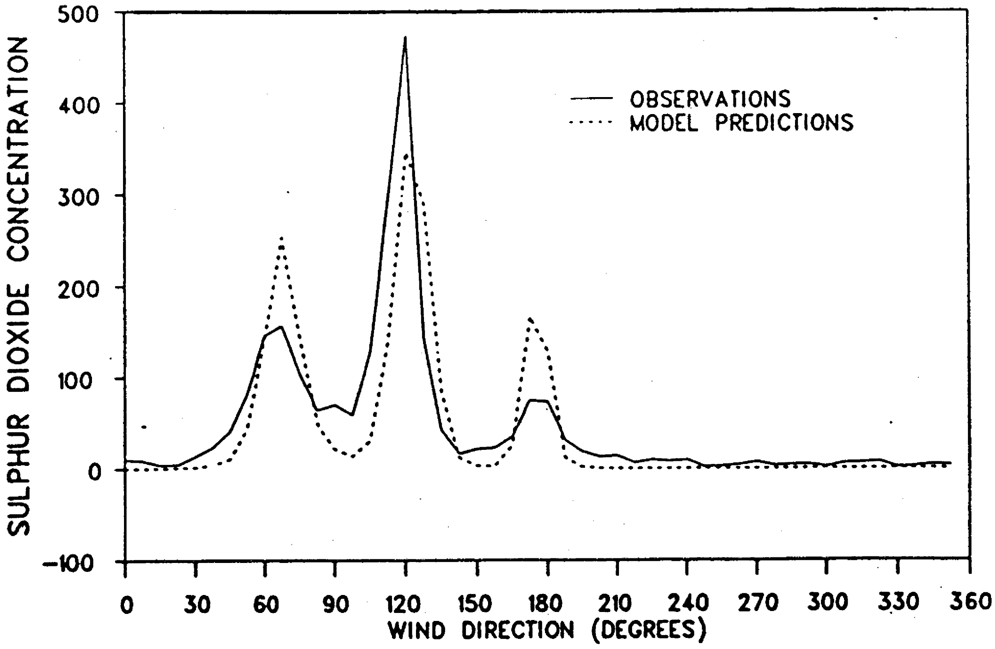

Figure 7.4: Same as Figure 7.3 except for the 1983 Technical School data set.

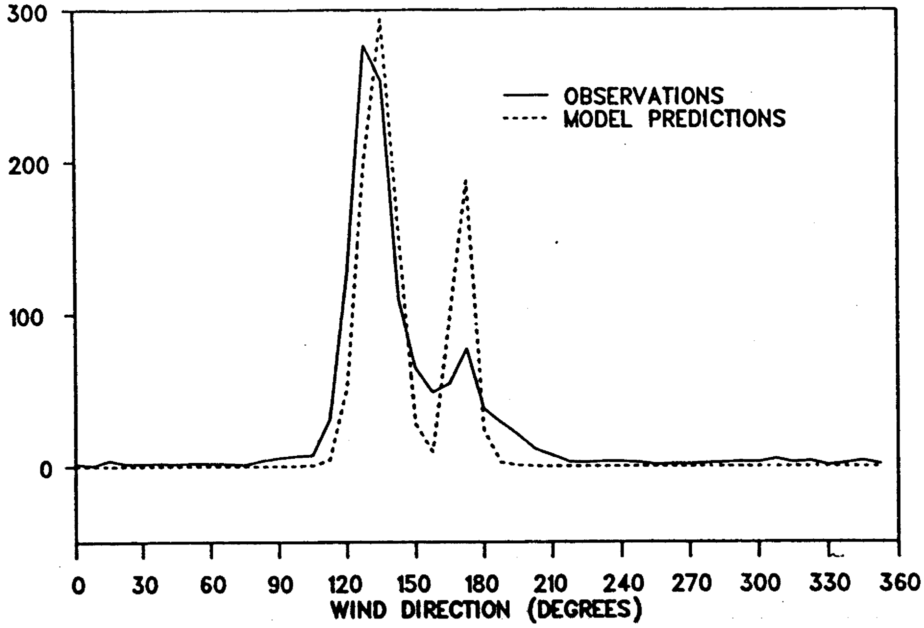

Figure 7.5: Same as Figure 7.3 except for the 1984 Hospital data set.

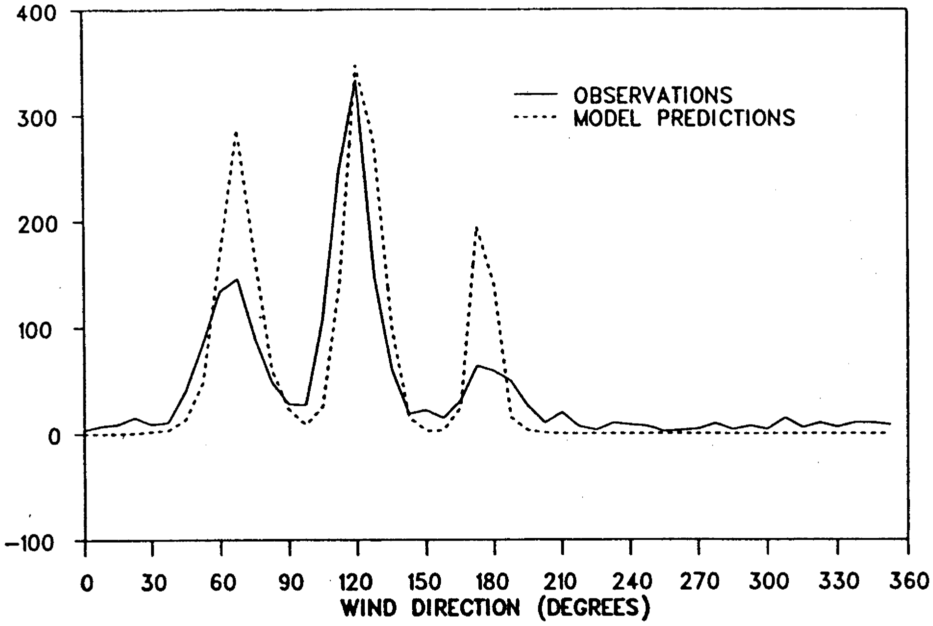

Figure 7.6: Same as Figure 7.3 except for the 1984 Technical School data set.

To test both temporally and spatially the calibration of the model derived from the Hospital data, three validation runs were made using the calibrated model but estimating the 1984 Hospital sulphur dioxide data and the 1983 and 1984 Technical School data. Figure 7.4 gives the variation of the observed and estimated concentrations with wind direction for the 1983 Technical School data set. The estimated concentrations were calibrated using the value obtained from the analysis of the Hospital data set. Three peaks in sulphur dioxide concentration are apparent in Figure 7.4. Again from Table 7.3 it may be noted that two of the sources share a similar bearing while the other two sources are at quite separate bearings. Figure 7.5 gives the variation in average sulphur dioxide concentration for the 1984 Hospital data set while Figure 7.6 presents a similar plot but for the 1984 Technical school data. Figures 7.3-7.6 illustrate that the Gaussian plume model, as developed in this chapter, provides a good explanation of the average variation of sulphur dioxide concentration with wind direction.

The diurnal variation of the average predicted and observed concentrations appear as Figure 7.7 for the 1983 Hospital data set while Figure 7.8 presents a similar plot for the 1983 Technical School data set. Figures 7.9 and 7.10 give the diurnal variation of predicted and observed sulphur dioxide concentrations for the 1984 Hospital and Technical School data sets respectively. These plots are typical of Gaussian plume model estimates in that the results are only reasonable for mean concentration conditions. The model does not provide a good explanation of the average pollutant concentrations during the morning hours 9-11 a.m. as these are most likely due to fumigation occurring as a result of the morning inversion layer breaking up. Clearly any further studies in the Kalgoorlie region should collect data relating to the mechanism producing the fumigation episodes. The results of this research could be incorporated into a more sophisticated deterministic model component.

The results of the calibration of the model predictions according to equation (7.8) are presented in Table 7.5 given previously. It shows that the correlation coefficient decreases with averaging time. Figure 7.11 illustrates the Gaussian plume model predictions plotted against the observed 24-h average data recorded at the Hospital monitor. Even though these model predictions yielded the highest value of the

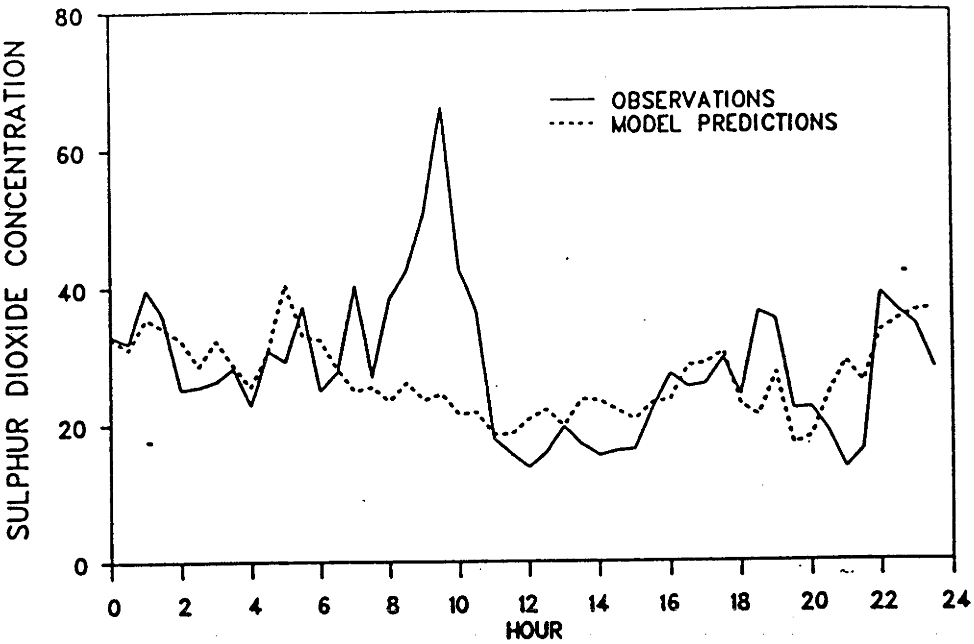

Figure 7.7: The diurnal variation of the average predicted and observed sulphur dioxide concentrations (μgm^{-3}) for the 1983 Hospital data.

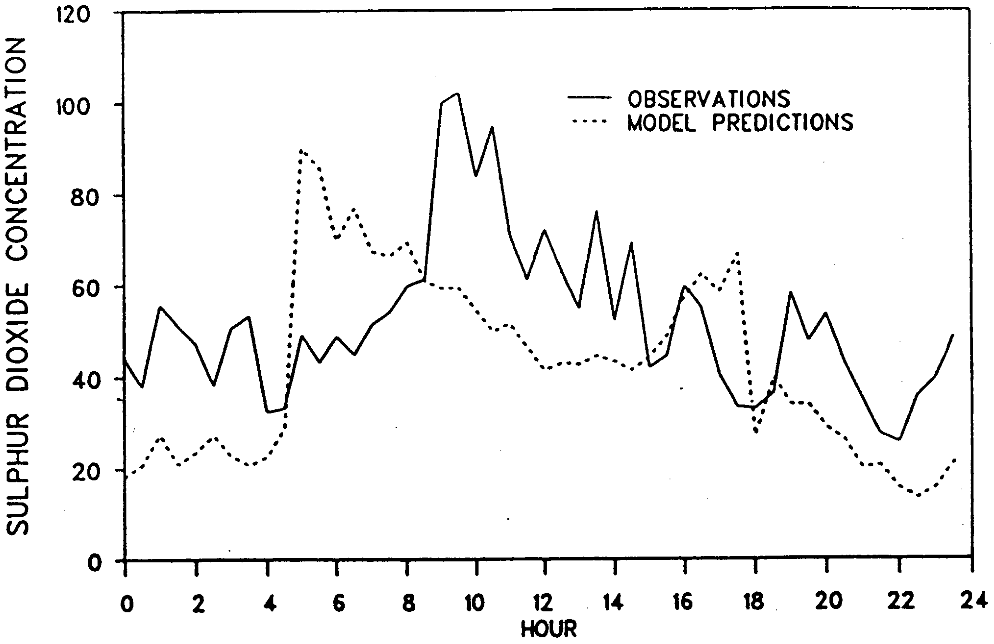

Figure 7.8: Same as Figure 7.7 except for the 1983 Technical School data.

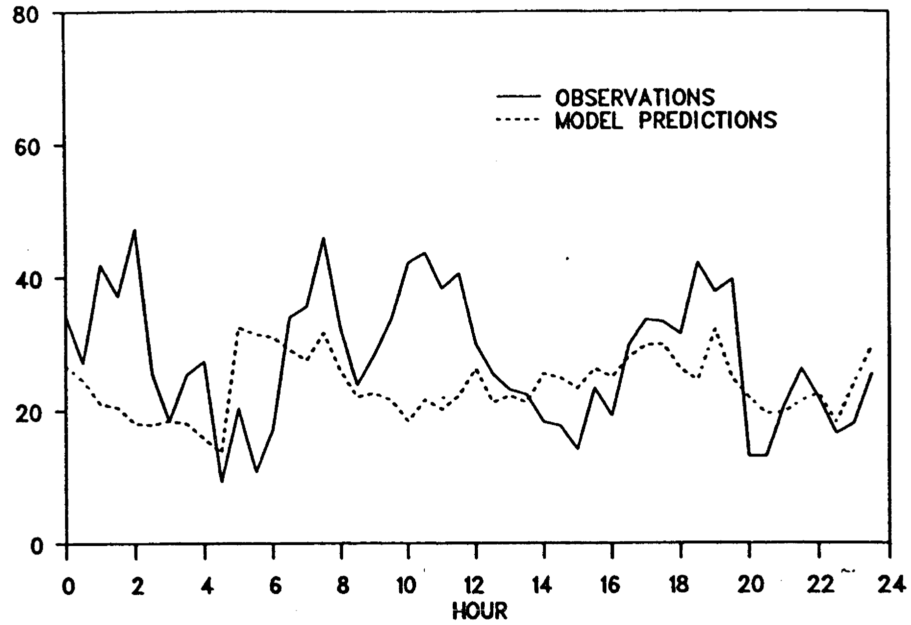

Figure 7.9: Same as Figure 7.7 except for the 1984 Hospital data.

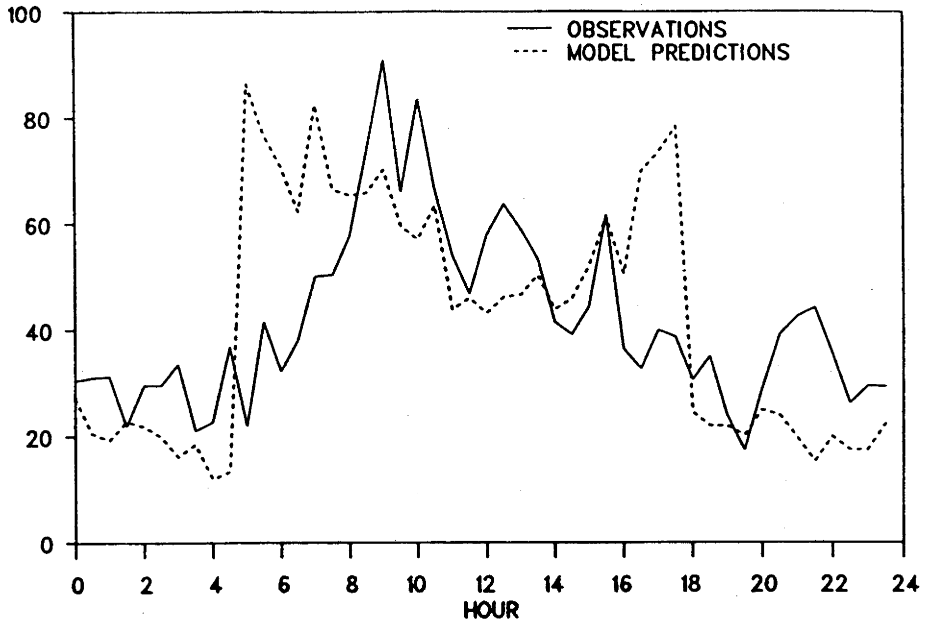

Figure 7.10: Same as Figure 7.7 except for the 1984 Technical School data.

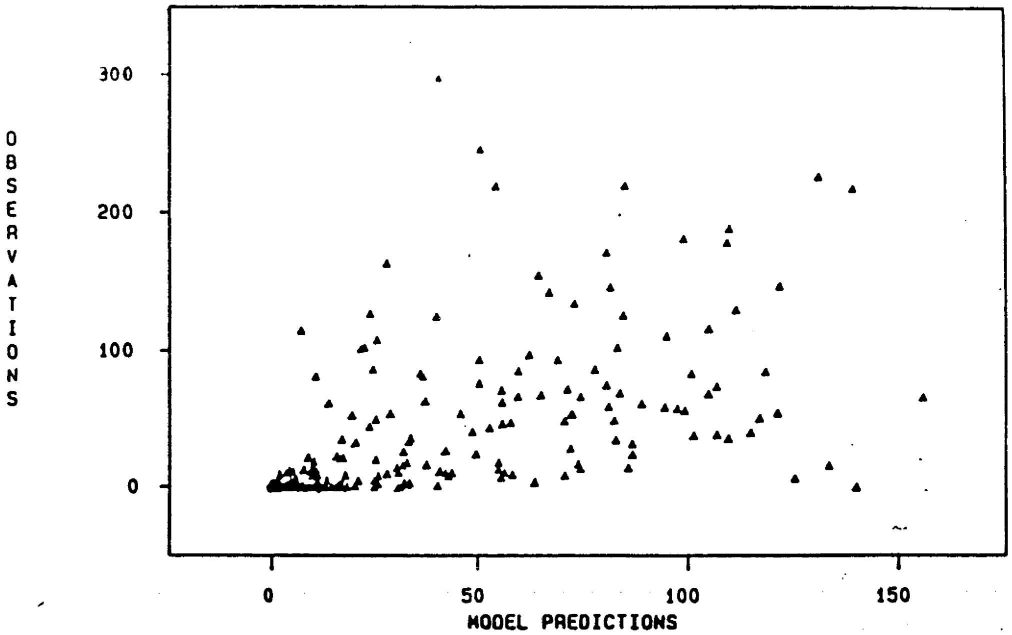

Figure 7.11 The predicted 24-h average sulphur dioxide concentrations plotted against the observations for the Kalgoorlie Base Hospital monitoring site.

correlation coefficient for the Hospital monitor data, the upper percentiles of the model predictions are in nearly all cases unrelated to the observations. This plot is considered typical of the results obtained when modelling these data.

Hanna et al. (1984) conducted a modelling study of dispersion from a point source using the Gaussian plume formula, modified for application in complex terrain, and obtained correlation coefficients between the predicted and observed concentrations for the two years of 24-h average non-zero data of 0.25. For 3-h average non-zero data a value of 0.24 was obtained. While these correlation coefficients are lower than observed here, as might be expected when modelling in complex terrain, Hanna et al. (1984) did find that the average model bias for the two highest concentrations at their 11 monitors were +7% for 3-h averages and +1% for the 24-h averages. While this result would indicate that the upper percentiles of the distribution were estimated well, the low correlation coefficients imply that this result may not always be obtained when estimating the upper percentiles of the distribution. Certainly their results agree with those here in that there is little correlation between model predictions of the high concentrations and observations.

In order to reduce the uncertainty associated with the direct prediction of the upper percentiles of the distribution using only the output of a deterministic model, the hybrid model employs the deterministic model output over the 50-90 percentile range only. Percentile ranges other than the 50-90 percentile range could be employed however, this percentile range was found to produce the highest values of the correlation coefficient when calibrating the model using the quantile-quantile comparisons with observations. The results of model calibration using equation (7.9) were presented previously in Table 7.6. Using this calibrated output the parameters of the two-parameter gamma, Weibull and lognormal models were estimated by applying of the method of maximum likelihood. With these parameter estimates the upper percentiles of the distribution of concentration can be evaluated. For the Hospital monitor data set, the estimated 98-percentile (x_{98}), the second highest concentration (x_{max-1}) and the maximum concentration (x_{max}) for the hybrid models with lognormal, gamma and Weibull distribution model components. Along with these model estimates the approximate 95% confidence intervals are presented. For comparison, the observed concentrations and the Gaussian plume model output, calibrated according to equation (7.8), are also given in Table 7.8. Finally, the results of a simple hybrid model using the exponential distribution with the single parameter estimated as the mean of the calibrated deterministic model output are also listed. This hybrid model was included as only one parameter must be estimated requiring the relatively simple calibration of the deterministic model output as given in Jakeman and Simpson (1985).

The results in Table 7.8 for the calibrated Gaussian plume model applied alone show that the model has systematically underpredicted the upper percentiles. The underprediction increases with decreasing averaging time. Of all the models considered in this chapter the Gaussian plume model output provides the worst estimates of pollutant concentration. The simple hybrid model based upon an exponential model improves upon these estimates in nearly all cases but the improvement is

Table 7.8: The observed 98-percentile concentration (x_{98}), the second highest concentration (x_{max-1}) and maximum concentration (x_{max}) with the Gaussian plume model estimates and hybrid model estimates with lognormal, gamma, exponential and Weibull statistical model components for the 1983 Hospital monitor data, Kalgoorlie, Western Australia. Approximate 95% confidence intervals for the lognormal, gamma and Weibull hybrid models are also listed.

Model	x_{98}	x_{max-1}	x_{max}
		24-h average data	
observed	217	243	295
Gaussian	122	137	153
lognormal	199(307)475	256(419)700	314(540)959
gamma	147(227)306	179(273)367	208(314)421
Weibull	177(226)318	207(273)397	234(315)471
exponential	141	202	241
		8-h average data	
observed	396	598	626
Gaussian	174	249	256
lognormal	425(588)815	586(861)1294	699(1068)1678
gamma	325(464)600	417(590)758	473(668)857
Weibull	376(455)587	460(573)765	510(646)877
exponential	238	461	584
		3-h average data	
observed	855	1042	1127
Gaussian	223	377	392
lognormal	819(1058)1370	1385(1927)2740	1668(2387)3508
gamma	642(827)1006	926(1185)1434	1049(1331)1609
Weibull	709(828)1008	980(1188)1500	1088(1334)1705
exponential	304	761	886
		1-h average data	
observed	1265	2547	2667
Gaussian	261	601	782
lognormal	1361(1625)1945	3027(3895)5108	3649(4780)6406
gamma	1069(1259)1441	1789(2084)2375	1994(2313)2633
Weibull	1141(1275)1454	1840(2132)2520	2034(2371)2822
exponential	356	1159	1326
		0.5-h average data	
observed	1495	3817	4099
Gaussian	231	747	1518
lognormal	1809(2071)2375	4691(5744)7153	5624(6978)8806
gamma	1421(1609)1788	2564(2875)3181	2830(3164)3499
Weibull	1492(1625)1793	2605(2929)3336	2860(3230)3697
exponential	391	1482	1680

not as great as when the gamma or Weibull models are used. Where the more complex hybrid models cannot be readily applied, this model could be used to provide improved estimates of the upper percentiles of the distribution of pollutant concentration.

For the hybrid model employing the two-parameter lognormal distribution the estimated concentrations are all overpredicted. This overprediction increases with decreasing averaging time. That this model does not describe the observations well is not surprising as, in Table 7.4, both the Kolmogorov statistic and the log likelihood statistic indicated that this distributional model did not provide a good fit to the observations. Instead these test statistics preferred the gamma or Weibull distributional models. The Kolmogorov statistic does prefer the Weibull model for 15 of the 20 data sets. However the log likelihood statistic prefers the Weibull model for only 13 of the 20 data sets. On many occasions the log likelihood statistics are very similar in magnitude. Based on these results the models should yield comparable estimates, with the Weibull model preferred. The results of Table 7.8 demonstrate that the Weibull and gamma models provide better models for x_{98}, x_{max-1} and x_{max} than any other model examined in this chapter and in all cases these models predict these quantiles to well within a factor of 2. The Weibull model appears to systematically predict higher values for the upper percentiles than the gamma model.

Using the model developed with the 1983 Hospital monitor data, a model verification exercise was performed. The data sets chosen were recorded at the Technical School monitor site with only a short time period of overlap with the data recorded at the Hospital monitor during 1983, and using a data set recorded during 1984. A data set recorded during 1984 at the Hospital monitor was also included in the model validation exercise. It should be noted that the Gaussian plume model predicts the average pollutant concentration for the 1983 Technical School data set as $43.2\mu gm^{-3}$ which is well within a factor of 2 of the observed mean of $52.1\mu gm^{-3}$ which is itself a factor of 2 greater than that observed in the model calibration. For the 1984 Technical School data set the estimated mean was $41.9\mu gm^{-3}$ and while the observed mean was $41.7\mu gm^{-3}$. The estimated mean for the 1984 Hospital data set was $23.8\mu gm^{-3}$ which compares with an observed mean of $27.8\mu gm^{-3}$. These results demonstrate the applicability of the calibration factor for the mean concentration. The results of the prediction of the upper

Table 7.9: The observed 98-percentile concentration (x_{98}), the second highest concentration (x_{max-1}) and maximum concentration (x_{max}) with the Gaussian plume model estimates and hybrid model estimates with lognormal, gamma, exponential and Weibull statistical model components for the 1983 Technical School monitor data, Kalgoorlie, Western Australia. Approximate 95% confidence intervals for the lognormal, gamma and Weibull hybrid models are also listed.

Model	x_{98}	x_{max-1}	x_{max}
		24-h average data	
observed	391	560	575
Gaussian	196	207	286
lognormal	209(296)420	235(341)503	270(409)636
gamma	151(254)357	168(281)393	193(317)442
Weibull	199(244)327	215(267)364	235(297)414
exponential	219	313	374
		8-h average data	
observed	697	920	1000
Gaussian	240	377	402
lognormal	486(650)871	668(948)1371	789(1157)1742
gamma	372(526)675	482(670)855	543(755)960
Weibull	429(510)641	523(639)830	575(714)941
exponential	296	545	639
		3-h average data	
observed	914	1332	1369
Gaussian	289	452	462
lognormal	910(1113)1364	1393(1809)2389	1615(2143)2906
gamma	731(943)1143	1022(1298)1565	1134(1438)1731
Weibull	811(917)1673	1054(1229)1482	1145(1348)1640
exponential	392	918	1062
		1-h average data	
observed	1212	2292	2418
Gaussian	349	923	1198
lognormal	1571(1828)2131	3172(3936)4966	3731(4701)6041
gamma	1266(1494)1705	2047(2389)2713	2262(2631)2986
Weibull	1339(1473)1649	2033(2308)2666	2215(2529)2940
exponential	470	1395	1889
		0.5-h average data	
observed	1513	2929	2999
Gaussian	350	1542	2199
lognormal	1811(2094)2439	4536(5663)7162	5368(6798)8709
gamma	1340(1557)1761	2381(2732)3074	2615(2990)3362
Weibull	1397(1527)1690	2323(2618)2999	2515(2852)3288
exponential	535	1871	2114

Table 7.10: The observed 98-percentile concentration (x_{98}), the second highest concentration (x_{max-1}) and maximum concentration (x_{max}) with the Gaussian plume model estimates and hybrid model estimates with lognormal, gamma, exponential and Weibull statistical model components for the 1984 Technical School monitor data, Kalgoorlie, Western Australia. Approximate 95% confidence intervals for the lognormal, gamma and Weibull hybrid models are also listed.

Model	x_{98}	x_{max-1}	x_{max}
		24-h average data	
observed	232	278	410
Gaussian	69	71	77
lognormal	193(267)370	221(316)459	253(378)576
gamma	140(227)313	159(255)350	181(287)393
Weibull	176(213)278	193(236)315	210(262)356
exponential	197	256	308
		8-h average data	
observed	518	678	854
Gaussian	178	228	307
lognormal	487(662)901	704(1021)1512	844(1267)1956
gamma	372(517)656	490(674)856	557(763)964
Weibull	422(505)642	529(654)860	586(736)983
exponential	198	334	395
		3-h average data	
observed	757	1153	1206
Gaussian	352	560	1070
lognormal	816(999)1227	1264(1645)2180	1469(1955)2659
gamma	654(841)1018	920(1164)1401	1021(1290)1551
Weibull	722(816)957	944(1101)1330	1027(1209)1473
exponential	269	588	686
		1-h average data	
observed	1153	1961	3115
Gaussian	550	1200	3792
lognormal	1452(1693)1977	2982(3711)4695	3520(4448)5734
gamma	1157(1364)1556	1879(2192)2489	2080(2416)2741
Weibull	1216(1334)1501	1854(2107)2437	2022(2311)2689
exponential	358	1001	1149
		0.5-h average data	
observed	1502	2840	3361
Gaussian	700	2580	8076
lognormal	1952(2194)2468	4483(5338)6448	5247(6319)7720
gamma	1584(1798)1994	2738(3077)3399	3004(3366)3715
Weibull	1643(1770)1927	2664(2950)3303	2888(3209)3609
exponential	440	1439	1637

Table 7.11: The observed 98-percentile concentration (x_{98}), the second highest concentration (x_{max-1}) and maximum concentration (x_{max}) with the Gaussian plume model estimates and hybrid model estimates with lognormal, gamma, exponential and Weibull statistical model components for the 1984 Hospital monitor data, Kalgoorlie, Western Australia. Approximate 95% confidence intervals for the lognormal, gamma and Weibull hybrid models are also listed.

Model	x_{98}	x_{max-1}	x_{max}
		24-h average data	
observed	227	327	414
Gaussian	76	86	91
lognormal	164(266)433	201(342)594	247(446)831
gamma	119(195)270	139(226)312	165(262)361
Weibull	149(196)289	169(229)347	193(266)419
exponential	124	165	200
		8-h average data	
observed	411	531	545
Gaussian	152	170	173
lognormal	358(494)682	442(638)937	516(774)1189
gamma	270(410)546	326(487)646	369(549)727
Weibull	327(395)511	374(462)614	411(515)698
exponential	120	223	267
		3-h average data	
observed	754	1013	1164
Gaussian	271	346	383
lognormal	663(854)1102	955(1309)1826	1121(1580)2281
gamma	516(701)877	691(922)1148	773(1032)1282
Weibull	581(676)824	727(872)1097	796(966)1231
exponential	167	410	484
		1-h average data	
observed	1042	1943	2167
Gaussian	411	552	560
lognormal	1102(1315)1572	2040(2601)3375	2404(3122)4140
gamma	880(1071)1248	1357(1630)1890	1505(1803)2089
Weibull	942(1051)1201	1360(1569)1852	1485(1726)2052
exponential	229	738	852
		0.5-h average data	
observed	1369	2419	2523
Gaussian	560	888	932
lognormal	1461(1669)1909	3062(3715)4578	3582(4405)5510
gamma	1182(1370)1544	1951(2238)2512	2144(2455)2753
Weibull	1233(1341)1479	1890(2128)2419	2062(2320)2650
exponential	272	1039	1188

percentiles of the distribution of observed concentrations recorded during 1983 at the Technical School monitor are presented as Table 7.9. Table 7.10 presents these results for the model validation using the 1984 Technical School data set, while Table 7.11 gives the results of a similar analysis except for the 1984 Hospital data set. As in the model calibration exercise the Gaussian plume model underpredicts the upper percentiles, while the hybrid model employing the exponential distribution improves upon nearly all model estimates. The hybrid model with a lognormal distribution component overpredicts the concentrations for the time averages other than for the 24-h averages. The hybrid models based upon the gamma and Weibull distributional models yield the best results with the Weibull model preferred at the shorter time averages. In all cases these models predict the observed concentrations, x_{98}, x_{max-1} and x_{max} well within a factor of 2. Of course further application of the hybrid modelling approach will be necessary in order to determine the limitations of the approach and to verify its wider applicability.

All estimates of the uncertainty in model predictions have been derived assuming that the observations are not autocorrelated. However, as noted in Chapter 3, the sulphur dioxide data at the 1-h and 0.5-h averaging times exhibit significant autocorrelation. This implies that the estimates of model uncertainty derived at these averaging times represent the minimum level of uncertainty likely to be encountered. Examination of the results of the model calibration and model validation exercises show that the estimates of model uncertainty provide reasonable bounds upon model uncertainty. That this is the case may be due to the incorporation of this uncertainty, at least to some extent, within the information matrix used to derive estimates of model parameter variances. Further investigation of the effects of autocorrelation should produce more accurate estimates of model uncertainty. Ideally this might mean applying a simple empirical model relating autocorrelation to model uncertainty.

The results of tables 7.4, 7.8 and 7.9-7.11 show how the hybrid model may be used for prediction. Given the limited data set at one site (here the 1983 Hospital data set), a Gaussian plume model is calibrated for the 50-90 percentile range. Statistical tests indicate that the appropriate statistical model (see Table 7.4) to be used generally is the Weibull model. So using the Gaussian plume model calibrated on the 1983 Hospital data set, the model would predict the results during 1983 and

1984 for the Technical School monitor as given in Tables 7.9 and 7.10, and at the Hospital monitor as given in Table 7.11 corresponding to the Weibull distribution. Clearly the agreement between these model predictions and the observed concentrations is well within the accuracy of a factor of 2 normally expected for air quality modelling.

7.7 Conclusions

A hybrid model has been developed for predicting the distribution of pollutant concentrations observed about elevated point sources. The two-parameter Weibull distribution component of the hybrid model was selected from amongst the two-parameter Weibull, gamma, lognormal and one-parameter exponential model alternatives using a selection procedure based upon test statistics of the Kolmogorov type and evaluation of the maximum of the respective likelihood functions.

The Gaussian plume model in this chapter predicted the mean concentration to well within a factor of 2. However prediction of the upper percentiles by this approach alone was accurate only within a factor of 4. These results compare favourably with those obtained by Nieuwstadt (1980).

Gaussian plume model calibration using quantile-quantile comparisons was found to yield significantly improved calibration factors than those based upon matched pair analysis. This method of deterministic model calibration produces accurate estimates of the percentile range employed in the estimation of the parameters of the distributional model. This is indicated by the high correlation coefficients observed, even when account is taken of the fact that the comparison is between the respective quantiles, not matched pairs of data. It should be noted that the correlation coefficient could be used as a test statistic, or a modified two sample Kolmogorov test, with which the following hypothesis could be examined: that over a particular percentile range the ordered observations and model predictions are drawn from the same distribution. Such a test would to some extent be dependent upon the underlying distribution of the observations.

Of the hybrid models examined in this chapter the model employing the Weibull distribution component yielded the best model predictions. The model predictions of the upper percentiles in nearly all cases fell well within a factor of 2 and offer substantial improvements over the estimates provided by the Gaussian plume model applied alone. Importantly, the observed concentrations, in nearly all cases, fell within the approximate 95% confidence intervals generated by the statistical component of the hybrid model. These 95% confidence intervals are of course the minimum 95% confidence intervals to be expected as they are evaluated assuming that the data are distributed exactly as modelled. However the results obtained in this chapter indicate that these confidence intervals provide useful bounds in that nearly all the observations fall within these intervals. Clearly wide application of this modelling approach would be needed to confirm this result, although the results obtained here and in Chapter 5 for an area source hybrid model application and in Chapter 6 for the line source hybrid model, are encouraging.

CHAPTER 8

ASSESSING COMPLIANCE WITH AIR QUALITY CRITERIA USING STATISTICAL MODELS OF RESTRICTED DATA SETS

8.1 Introduction

In order to obtain a clear picture of the air quality within an airshed of interest it is necessary to have as many air pollution monitors operating as possible. Unfortunately the cost of installing and operating air quality monitors will limit their use. The problem then is to make the network as 'representative' as possible both in space and time. Preferrably all 'hot spots' and all land use types such as urban, residential and industrial should be included and the sampling frequency at each site should allow compliance with chosen air quality standards ~~criteria~~ to be reliably assessed. Clearly, continuous monitoring achieves the latter requirement.

Usually most monitors are fixed in position and record continuously or intermittently (e.g. total suspended particulate high volume samples collected every six days). A more cost effective method may be to have mobile monitors which can be used at more sites than fixed monitors and can be operated at a similar expense to that of a few continuous monitors. The problem then becomes one of obtaining a representative sample. It is reasonably straightforward to design a random sampling strategy to yield reasonable estimates of annual mean concentrations (e.g. see Ott and Mage, 1981; Simpson, 1984). However it is a different problem when it comes to obtaining reasonable estimates of a cumulative frequency distribution from a limited sample so that exceedances of 98-percentile or maximum air quality standards can be estimated. In this chapter statistical models are developed to fulfil that need. The models constructed employ a mixture of fixed continuous air quality monitors and intermittent monitors which can be mobile or fixed, allowing the most effective spatial representation of pollutant levels to be obtained within the prevailing economic constraints.

The first model, the simplest and most successful, is an empirical quantile-quantile model. The model predicts the upper percentiles of a restricted data set, provided a more complete data set from another site in the same airshed and collected over the same time period is available. Clearly, any modelling of a restricted data set from a base data set will only be as good as the information in the latter. Ideally, the base data set should be able to provide a reliable sample of air quality observations for comparison with chosen air quality standards ~~criteria.~~ Continuous recordings will obviously provide this. However, sufficiently regular or a sufficient number of random recordings would also achieve this goal. The problem of what constitutes a representative base data set is not investigated here. The base data set employed records 24-h acid gas levels five days per week which is the most complete data set available. An empirical quantile-quantile model is derived from quantiles of both the more complete and restricted data sets which then can be used to predict the upper percentiles of the restricted data set (not observed because of the limited sampling). The model assumes both the restricted and more complete data sets are of the same distributional form, an assumption that can be tested by the two-sided Kolmogorov-Smirnov two-sample test.

The second model assumes a distributional form for the model, the parameters of which can be estimated from the restricted data set and the distribution then used to estimate the upper percentiles. Four standard distributions are used here for that purpose: the two-parameter gamma, lognormal, and Weibull and the one-parameter exponential. The third model is an improvement on the second one in that, for each data set, the 'best' distribution is selected from amongst the four chosen here using a goodness-of-fit test. Here the test developed in Chapter 4 is used. Clearly this 'best fit' method should be an improvement on the second. The best test of the empirical quantile-quantile model is to compare its results with those of the third model.

For hourly carbon monoxide observations, Ott and Mage (1981) were able to demonstrate that accurate estimates of the arithmetic mean, and associated 95% confidence limits, could be obtained by random sampling. Their approach was based on the application of the Central Limit Theorem and thus remains independent of the nature of the original distribution. However, maximum or near maximum concentrations, in terms of which many air quality standards are written, may not be similarly estimated. Ott and Mage (1981) originally proposed that 'random' sampling could take place between the hours of 9 a.m. and 5 p.m. However, Simpson (1984) identified the importance of both the diurnal and seasonal variations observed in ambient air quality and found that sampling of hourly pollutant data should be carried out as a completely random process to obtain best estimates of the mean concentration for ozone, nitrogen dioxide and total suspended particulates. Simpson (1984) also found that continuous recordings 1 week out of 4 yielded good results.

In this chapter the various model performances are examined using restricted data sets which have been generated using a sampling strategies of 1 day in 4, 1 in 6, 1 in 8, and 1 in 12. Such strategies should satisfactorily account for seasonal variations. Since the models are developed with 24-h average acid gas concentrations, problems associated with the diurnal variation of pollutant concentrations will not be present.

8.2 The data set

The acid gas data set used consists of daily levels collected by the Health Division of the Newcastle City Council at their Watt Street, City and Mounter Street monitoring stations in Newcastle, Australia, over a 10 year period from January 1972 to December 1981; at the Turton Road monitoring station over a 9 year period from January 1973 to December 1981; and at their Seaview and Elder Street monitoring stations over an 8 year period from January 1974 to December 1981. The acid gas levels were determined by scrubbing ambient air at a constant rate through a dilute solution of hydrogen peroxide and sulphuric acid. Thus any sulphur dioxide present in the sample is converted to sulphuric acid. The resulting increase in acidity due to this or other acid gases is determined by titration. The method is based on the British Standard Method No. 1747 Part 3. The 24-h average readings are taken daily commencing at 9 a.m. for five days per week.

The locations of the monitoring sites relative to the industrial area emitting acid gases are illustrated in Figure 5.1. It should be noted that the Mounter Street monitor lies in closest proximity to the industrial area. The Watt Street and City monitors are located within the central business district. The Seaview Street, Turton Road and Elder Street monitors are situated in the surrounding urban area. In this analysis no distinction between different land use categories is drawn. This does lead to problems and indications as to how future work may improve on the results are presented here.

8.3 The empirical quantile-quantile model

We denote ~~Denoting~~ two data sets as x_i, (i = 1 to n) and y_j, (j = 1 to m) and their empirical cumulative distribution functions as Q_x and Q_y respectively. Then an empirical quantile-quantile plot is a plot of $Q_y(p)$ against $Q_x(p)$ for a range of p-probability values where p may vary from 0 to 1 (Wilk and Gnanadesikan, 1968). If the two distributions were identical all the points would fall on the line y = x. This simple model relating the quantiles of the two data sets may be applied only where both data sets have been drawn from the same distribution. Chambers et al. (1983) have extended this simple model to allow the quantiles to differ by both an additive and multiplicative constant. Thus the two data sets would have the approximate relationship

$$Q_y(p_i) = \alpha\, Q_x(p_i) + \beta \qquad (8.1)$$

where the parameters α and β may be estimated using ordinary least squares. Following the terminology of Chambers et al. (1983) equation (8.1) is referred as the empirical quantile-quantile model.

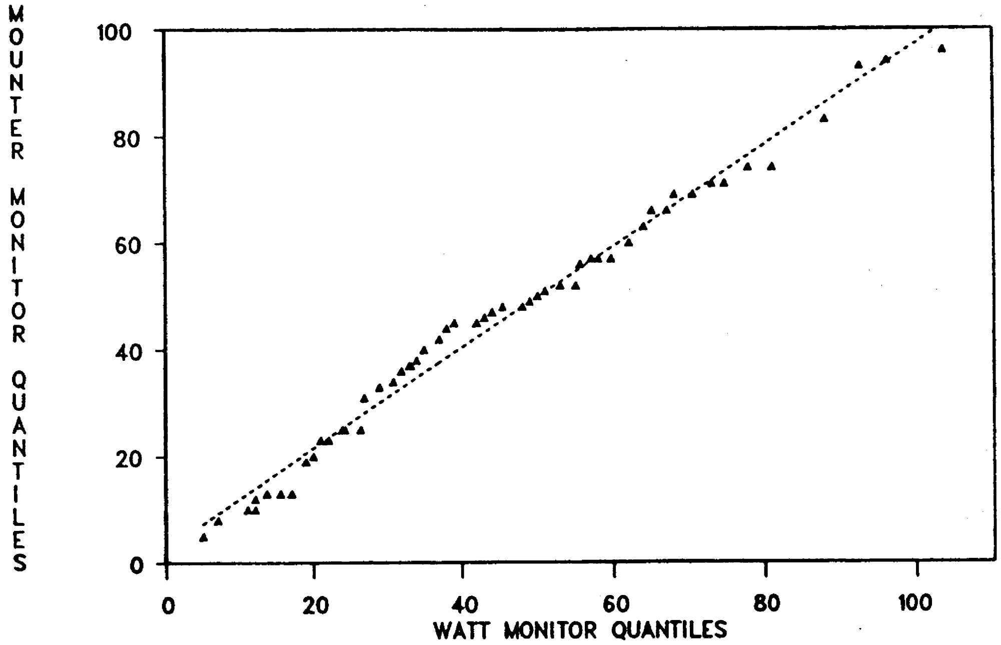

Figure 8.1: Acid gas concentrations ($\mu g\ m^{-3}$) selected from the Watt Street and Mounter Street monitors data sets with the fit according to equation (8.1).

With data sets of equal size an empirical quantile-quantile plot consists of a plot of the sorted data values paired from the lowest to the highest values for each data set. However, here one data set is larger than the other. The usual practice is to employ all the sorted values in the smaller data set and to interpolate a corresponding set of quantiles from the larger set (Chambers et al., 1983). In order to determine the corresponding quantile in the larger data set the percentile at which each of the smaller data set values fall must be estimated. Many empirical distribution functions are available for this purpose (Looney and Gulledge, 1985). The form of the empirical distribution function applied is that suggested by Chambers et al. (1983) which, for a sample of size n, is

$$p_i = (i - 0.5)/n \qquad (8.2)$$

Now suppose that y_j is the smaller data set and x_i is the larger then y_j, which is the $(j - 0.5)/m$ quantile of the y data, is matched with the interpolated $(j - 0.5)/m$ quantile of the x data set. Thus the required order statistic in the larger data set is determined from

$$v = \frac{n}{m}(j - 0.5) + 0.5 \qquad (8.3)$$

If v is not an integer this value is separated into an integer component i and a fractional component θ. The interpolated quantile is evaluated as

$$Q_x((j - .5)/m) = (1 - \theta)x_i + \theta x_{i+1} \qquad (8.4)$$

An example of an empirical quantile-quantile plot using the acid gas data recorded at the Watt and Mounter monitors is presented as Figure 8.1. The least squares fit of equation (8.1) to these data yielded parameter estimates and associated standard errors of $\alpha = 0.946 \pm 0.015$ and $\beta = 2.602 \pm 0.750$ with a correlation coefficient of 0.994.

8.4 Kolmogorov-Smirnov two-sample test

Given independent random samples of sizes n, m respectively from continuous distribution functions $F_n(x)$ and $F_m(x)$, the hypothesis following is tested

$$H_o: F_n(x) = F_m(x), \text{ all } x \qquad (8.5)$$

This class of nonparametric problem can be solved by distribution free methods which do not depend on the form of the underlying distributions at all, provided that they are continuous (Kendall and Stuart, 1973). The two-sided Kolmogorov-Smirnov two sample test criterion, denoted by $D_{n,m}$ is the maximum absolute difference between the two empirical distributions, $S_n(x)$ and $S_m(x)$, of $F_n(x)$ and $F_m(x)$ respectively

$$D_{n,m} = \max \mid S_n(x) - S_m(x) \mid \qquad (8.6)$$

Tables of the quantiles of this test statistic are readily available for small samples $(n < 40)$ with approximations for larger samples (Birnbaum and Hall, 1960). The Kolmogorov-Smirnov test is sensitive to all types of differences between the cumulative distribution functions (Gibbons, 1971; Kendall and Stuart, 1973). The data sets tested here are the restricted data set and that extracted from the complete data set modified according to equation (8.1).

An alternative test to the Kolmogorov-Smirnov test, the Cramer-von Mises test for two samples, could be applied to the data sets examined here. However, Conover (1980) notes that there is probably little difference in power between the two tests while the Kolmogorov-Smirnov test can be evaluated more easily. Thus in order to demonstrate the efficacy of the empirical quantile-quantile model the Kolmogorov-Smirnov test was applied.

8.5 Application of the models

Of the six monitoring sites the Watt Street, Mounter Street and City data sets span the longest time interval (10 years). Thus the data sets at these monitoring sites have been chosen, each in turn, to represent the complete data sets. Three sites within the Newcastle airshed have been chosen in order to illustrate how the empirical quantile-quantile model performance is affected by the selection of the monitoring site representing the complete data set.

Having decided on one of the City, Watt Street and Mounter Street sites as the base data set (providing 10 years of continuous data), then restricted data sets are made up from one of the other five sites (see Section 8.2). There are 45 years of continuous data therefore which can be chosen as restricted data sets. Restricted data sets were constructed in four ways: choosing 1 day out of every 4, 1 out of 6, 1 out of 8 and 1 out of 12. All the data sets were chosen to be of 1 year duration thereby providing 10 complete sets for each of the base stations. For each year, the quantiles for the restricted data sets were constructed from those of the complete data set using equations (8.3) and (8.4). Given these quantiles and those of the complete data set, the parameters (α,β) of the quantile-quantile model in equation (8.1) were estimated using ordinary least squares analysis. Using these parameter estimates and the upper percentiles of the complete data set, the corresponding upper percentiles of each restricted data set were then estimated.

For each of the restricted data sets the parameters of the two-parameter lognormal, gamma, Weibull and one-parameter exponential distributions were estimated using the method of maximum likelihood (see Chapter 3). The model considered to have the 'best fit' was identified from amongst these four distributions using the procedure developed in Chapter 4. The identification of the 'best fit' distributional models ~~were~~ was based upon the analysis of the limited data sets only.

Estimates of the upper percentiles of the restricted data sets were then calculated using two statistical models. In the first, a distributional form was assumed with the estimated maximum likelihood parameters used to produce estimates of the upper percentiles. The four standard distributional forms mentioned previously were each chosen in turn. In the second model, for each data set only the 'best fit' distributional form was used to estimate the upper percentiles.

Model performance was evaluated by examining the ability of each model to estimate the upper percentiles of the distribution of pollutant concentrations. This region has been selected as many regulatory standards refer to these upper percentiles. In particular we examine for each model the prediction of the three highest concentrations, denoted respectively as x_{max}, x_{max-1} and x_{max-2}, and the 98-percentile, x_{98}. This last percentile was chosen as the World Health Organization goal for 24-h average acid gas concentrations refers to this percentile.

At each of the percentiles listed above the agreement between observed (x_o) and predicted percentiles (x_p) was assessed using a measure of the average relative bias of the estimates and the average relative root mean square error (rmse). These criteria were recommended by Fox (1981) for the assessment of air quality models and are defined as follows

$$\text{bias} = \frac{1}{N} \sum_{i=1}^{n} (x_p - x_o)/x_o \tag{8.7}$$

$$\text{rmse} = \{\frac{1}{N} \sum_{i=1}^{n} ((x_p - x_o)/x_o)^2\}^{0.5} \tag{8.8}$$

where N is the number of acid gas data sets.

8.6 Results and discussion

For each of the concentrations x_{max}, x_{max-1}, x_{max-2} and x_{98}, the average relative root mean square errors for each of the models considered are presented as Table 8.1 for the case of restricted data based on a choice of one day in four sampling. The average relative biases for these same data sets are presented in Table 8.2. The results are presented for the three alternative cases where the Watt Street, Mounter Street and City monitors represent the complete data set. These average values were determined as a result of the analysis of 45 acid gas data sets based on acid gas observations from six monitoring sites for periods of 8-10 years.

Consider first the performance of the second approach where four distributional models have been applied separately to all 45 data sets. It should be noted that the gamma model yields the lowest rmse values overall. This result is not surprising given that in Chapter 5 for these data the gamma distribution was preferred by the majority of data sets when compared with the lognormal model. The lognormal model does appear to provide slightly better estimates of the x_{max} and x_{max-1} concentrations however this model performs poorly at the x_{max-2} and x_{98} concentrations when compared with the gamma model. The exponential and Weibull models both perform badly in comparison with the gamma distributional model. The results of Table 8.1 are supported by the average relative bias results as listed in Table 8.2. These data show that the exponential and lognormal models consistently overpredict the upper percentiles while the gamma and Weibull models underpredict. In particular the lognormal model overpredicts the maxima.

There may be errors however, in assuming one distributional form for all the years of data and at all sites. Certainly it would appear that the gamma distribution is preferable but a more accurate procedure is to identify the 'best' distribution to fit the data for individual sites and years using a goodness-of-fit test. This procedure has been adopted here and comprises the third statistical approach. The model identification procedure has been applied to the restricted data sets to select the 'best fit' model for each restricted data set from amongst the four distributional models which had previously been employed separately. Again parameter estimates were those derived using the method of maximum likelihood. The goodness-of-fit method used is described in Chapter 4.

A comparison of the 'best fit' model results with any of the four distributional models applied separately indicates that a significant improvement in the rmse has been obtained (listed in Table 8.1). The one exception involves the City monitor whose x_{max} estimates indicate only small improvement. The results of Table 8.2 further confirm the improved performance of the 'best fit' model. Very much smaller values of

Table 8.1: Average relative root mean square errors for the empirical quantile-quantile, exponential, lognormal, gamma, Weibull and best fit models where the Watt Street, Mounter Street and City monitors are the complete data sets.

Model	x_{max}	x_{max-1}	x_{max-2}	x_{98}
		Watt Street monitor		
quantile-quantile	.228	.186	.157	.127
exponential	.393	.381	.364	.351
lognormal	.348	.307	.245	.216
gamma	.369	.334	.233	.172
Weibull	.449	.372	.249	.170
best fit	.336	.295	.192	.146
		Mounter Street monitor		
quantile-quantile	.411	.197	.148	.136
exponential	.418	.390	.372	.356
lognormal	.348	.308	.250	.220
gamma	.296	.304	.222	.172
Weibull	.321	.313	.224	.166
best fit	.201	.217	.162	.145
		City monitor		
quantile-quantile	.365	.203	.167	.118
exponential	.420	.397	.384	.369
lognormal	.345	.307	.253	.220
gamma	.347	.333	.229	.170
Weibull	.433	.372	.245	.166
best fit	.333	.294	.188	.140

Table 8.2: Average relative bias for the empirical quantile-quantile, exponential, lognormal, gamma, Weibull and best fit models where the Watt Street, Mounter Street and City monitors are the complete data sets.

Model	x_{max}	x_{max-1}	x_{max-2}	x_{98}
		Watt Street monitor		
quantile-quantile	.080	.034	.057	.047
exponential	.339	.317	.316	.316
lognormal	.205	.158	.150	.143
gamma	-.155	-.123	-.086	-.044
Weibull	-.279	-.207	-.151	-.088
best fit	-.089	-.079	-.051	-.020
		Mounter Street monitor		
quantile-quantile	-.214	-.043	.009	.046
exponential	.371	.331	.324	.319
lognormal	.251	.175	.161	.147
gamma	-.089	-.096	-.070	-.037
Weibull	-.192	-.167	-.125	-.075
best fit	-.006	-.037	-.025	-.007
		City monitor		
quantile-quantile	-.069	-.019	-.027	-.020
exponential	.375	.339	.344	.338
lognormal	.214	.151	.153	.143
gamma	-.131	-.123	-.076	-.038
Weibull	-.267	-.217	-.149	-.089
best fit	-.094	-.096	-.056	-.025

the relative bias have been obtained using this model. Thus the systematic overprediction or underprediction of pollutant concentrations at the percentiles of interest has been reduced. This should, in turn, produce a concomittant improvement in the decisions regarding the level of control required to meet air quality standards both in the short and long terms. The results of Table 8.1 and Table 8.2 demonstrate the importance of the application of a model identification procedure.

Having examined the performance of the distributional models applied to the restricted data sets, the empirical quantile-quantile model performance is examined. Again, it should be noted that this model cannot be applied to a restricted data set alone. The empirical quantile-quantile model assumes that at one monitoring site within the airshed, a complete or near complete record is available. The important advantage of this method is that the identification of the best distributional form that should be applied to a particular data set is not necessary. The model also avoids the more computationally demanding evaluation of the parameters of the identified distributional form since the empirical quantile-quantile model parameters may be easily estimated using ordinary least squares analysis. The only assumption necessary for the application of the empirical quantile-quantile model is that the two data sets are drawn from the same distributional form which need not be specified.

In order to verify that this assumption holds, the two-sided Kolmogorov-Smirnov two-sample test was applied. Here the maximum absolute difference between the two empirical distribution functions, evaluated according to equation (8.6), is compared with tables of the test statistic (Birnbaum and Hall, 1960). This analysis found that the hypothesis that the two data sets are drawn from the same distribution, as stated in equation (8.5), could be rejected at the 95% confidence level on only one occasion out of a total of 135 tests. This was where the City monitor data were considered as the complete data set. This result implies that the assumption of similar though unspecified distributional form for data recorded at different sites within the Newcastle airshed, when related according to equation (8.1), is reasonable. This conclusion is borne out in the empirical quantile-quantile model results.

Table 8.1 demonstrates that the empirical quantile-quantile model produces the lowest values of rmse in nearly all cases. The only exceptions are for the maximum concentrations when the Mounter Street and City monitor data sets are considered the complete data sets. For the City data set the rmse value is only slightly greater than the 'best fit' model values. The average relative bias results of Table 8.2 are further evidence that the empirical quantile-quantile model compares favourably with the 'best fit' model. On many occasions the empirical quantile-quantile model produces lower values of the average relative bias than the 'best fit' model. Only in one case, for the estimates of x_{max} concentration when the Mounter Street data are the complete data sets, does the empirical quantile-quantile model produce a much larger average relative bias than the 'best fit' model.

The results for Mounter Street can be expected to be the worst. As stated previously, the Mounter Street monitor is sited in the industrial area while the other monitors lie in urban or residential areas. It is therefore possible that the empirical model results could be improved if monitor data were only paired with those from the same land use category in applying the quantile-quantile model. However the City and Watt Street sites, being city based, are also not exactly the same as the other three sites so clearly the dependence on land use is not strong here.

Figures 8.2-8.4 present the observed, 'best fit' model estimates and the empirical quantile-quantile estimates of x_{98} for the case where the Watt Street, Mounter Street and City monitors are respectively considered as the complete data sets. These plots indicate the excellent performance of the empirical quantile-quantile model regardless of the monitoring site chosen as the complete data set and 'best fit' models when estimating the 98-percentile concentration. Figures 8.5-8.7 present similar plots to that of Figures 8.3-8.5 except for x_{max}. Again the predictive ability of both the empirical quantile-quantile model and 'best fit' models are similar. As would be expected from the results of Tables 8.1 and 8.2, increased uncertainty associated with model predictions of x_{max} over those obtained at the 98-percentile has resulted. An increased variance between the empirical quantile-quantile and 'best fit' model estimates is also apparent.

The other monitoring strategies that have also been considered require a 1 day in 6, a 1 day in 8 and a 1 day in 12 monitor operation

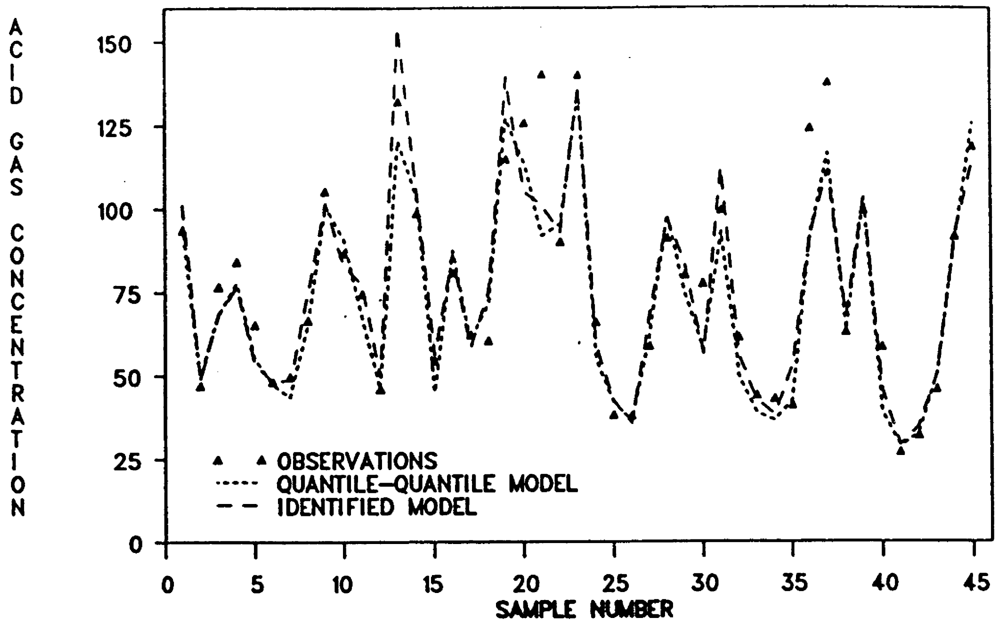

Figure 8.2: The 98-percentile acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with Watt Street monitor data as the complete data set.

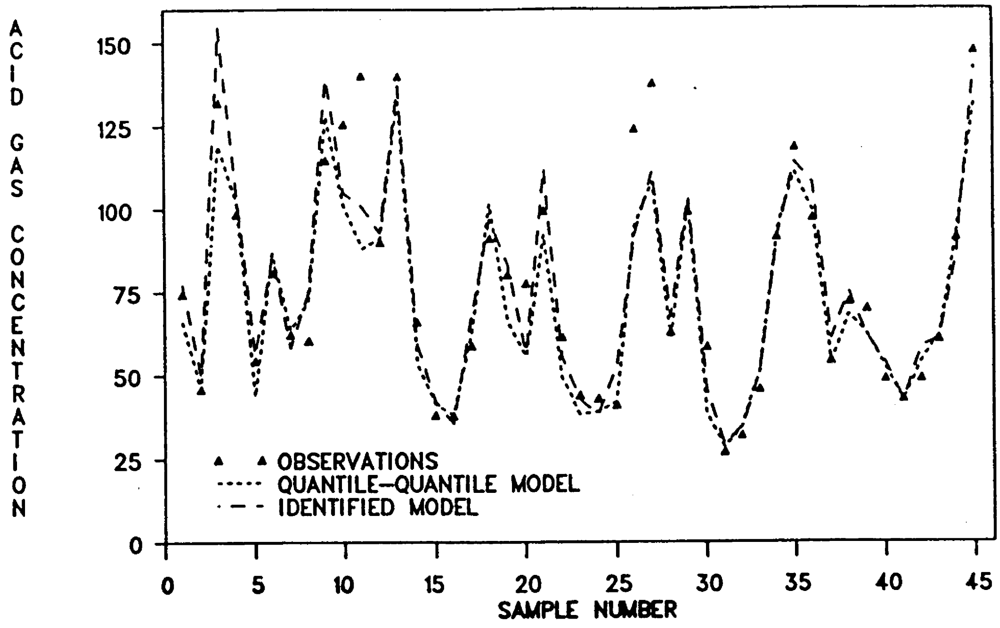

Figure 8.3: The 98-percentile acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with the Mounter Street monitor data as the complete data set.

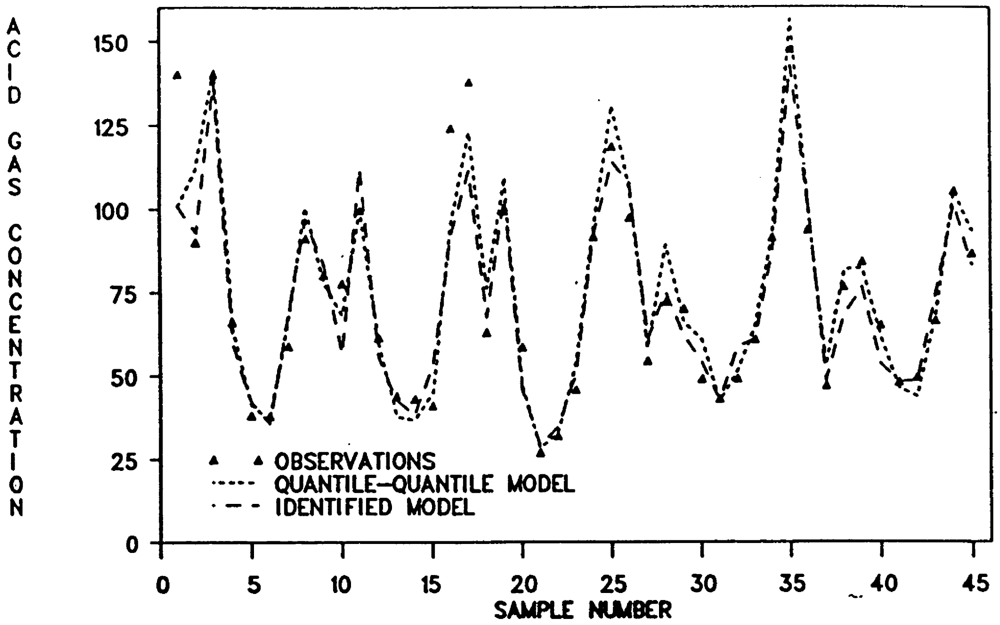

Figure 8.4: The 98-percentile acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with the City monitor data as the complete data set.

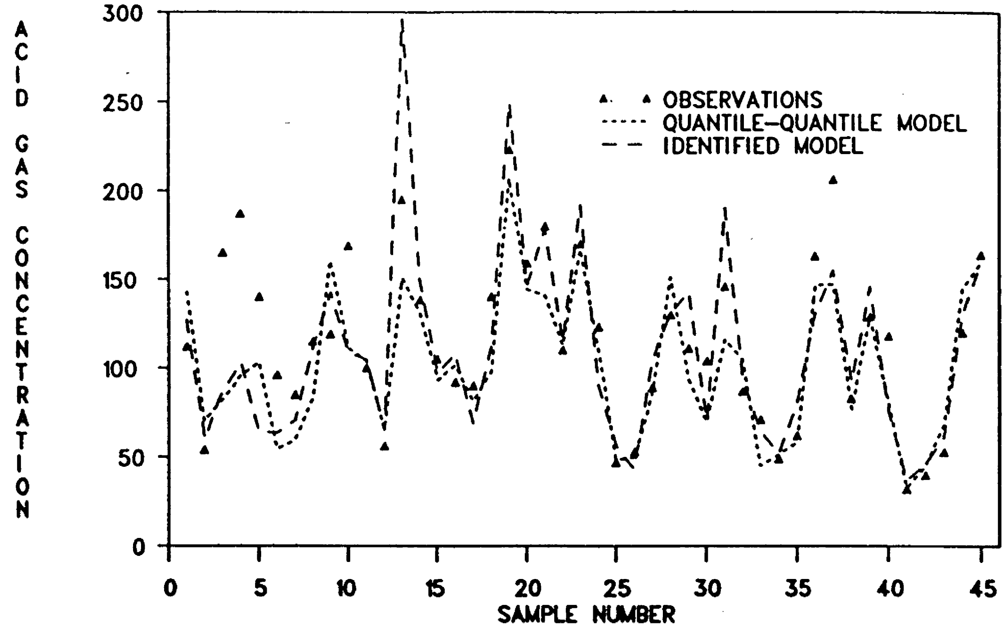

Figure 8.5: The maximum acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with Watt Street monitor data as the complete data set.

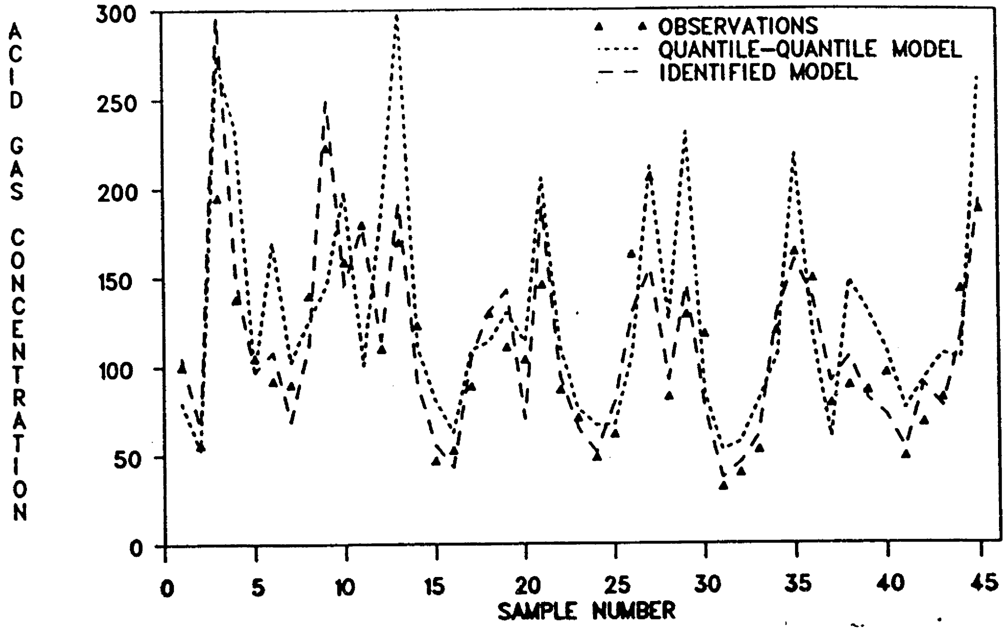

Figure 8.6: The maximum acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with the Mounter Street monitor data as the complete data set.

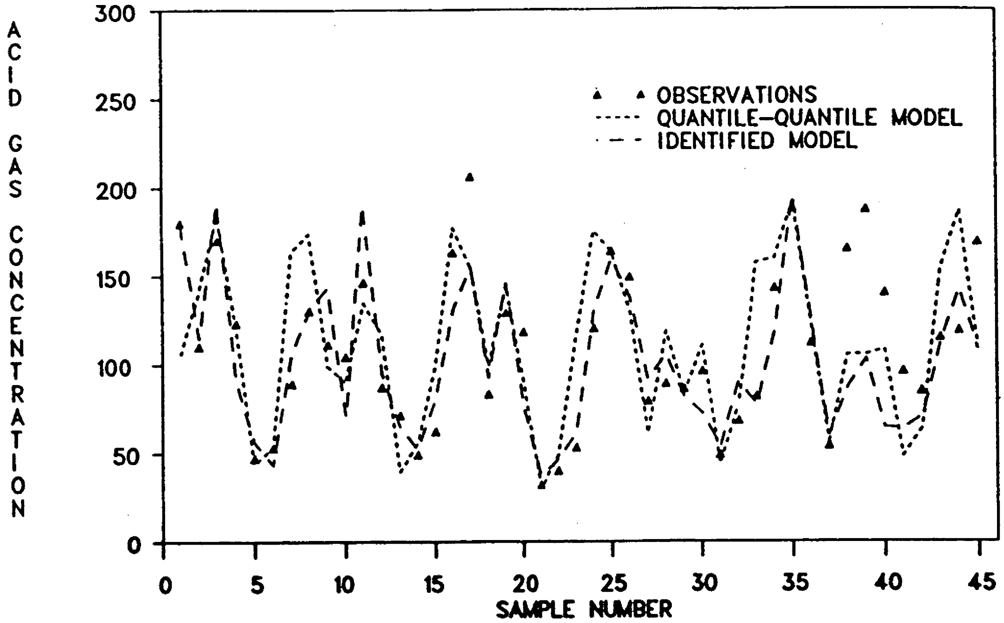

Figure 8.7: The maximum acid gas concentrations ($\mu g\ m^{-3}$) with the empirical quantile-quantile model and identified model predictions with the City monitor data as the complete data set.

(strategies of 1 day in 7 should be avoided as this would mean collecting samples on the same day of the week throughout the year). As the application of a distributional model a priori does not produce the best results, at least for the restricted data sets examined here, only the performance of the empirical quantile-quantile and 'best fit' models are considered.

The results of applying the empirical quantile-quantile and best fit models to restricted data sets based upon 1 day in 6, 1 day in 8 and 1 day in 12 monitoring strategies are given in Tables 8.3 and 8.4. Table 8.3 lists the rmse values while Table 8.4 presents the average relative bias results. The results of Table 8.3 demonstrate, in terms of rmse, that the empirical quantile-quantile model provides the best model. Only for the Mounter Street monitor x_{max} estimates does the 'best fit' model yield better rmse results. That the empirical quantile-quantile model is the better model is supported by the average relative bias results of Table 8.4. For 24 of the 36 measures of bias the empirical quantile-quantile model results are lowest. Only for the Mounter Street monitor estimates of x_{max} does the empirical quantile-quantile model perform significantly worse than the 'best fit' model. Again these results suggest that pairing monitored data sets with those from a similar land use category may improve these results.

The cause of the decreased performance of the 'best fit' model with the reduction in the sample size of the restricted data set can be attributed to the increased uncertainty of the parameter estimates of the identified distributional model. The poorer performance can also be attributed to the increased difficulty in identifying the best distributional model from the smaller data sets.

The results in Tables 8.1 and 8.3 also indicate the problems with intermittent sampling. The rmse values generally increase as the data set decreases, especially for the 'best fit' case. However Tables 8.2 and 8.4 indicate no such trend for the relative bias. It would be expected that as the data sample decreases the model fit to observations would worsen. The rmse values show this behaviour, but not the relative bias. Hence it would appear that the rmse factor is the more sensitive goodness-of-fit indicator. It would also appear that the relative accuracy of the quantile-quantile model compared to the 'best fit' model

Table 8.3: Average relative root mean square errors for the empirical quantile-quantile and 'best fit' models where the Watt Street, Mounter Street and City monitors are the complete data sets for 3 monitoring strategies, 1 day in 6, 1 day in 8 and 1 day in 12.

Model	x_{max}	x_{max-1}	x_{max-2}	x_{98}
		Sampling strategy of 1 day in 6		
		Watt Street monitor		
quantile-quantile	.229	.193	.161	.146
best fit	.390	.308	.218	.181
		Mounter Street monitor		
quantile-quantile	.411	.208	.146	.151
best fit	.336	.272	.204	.184
		City monitor		
quantile-quantile	.357	.207	.181	.150
best fit	.391	.331	.226	.183
		Sampling strategy of 1 day in 8		
		Watt Street monitor		
quantile-quantile	.244	.233	.205	.177
best fit	.468	.447	.324	.250
		Mounter Street monitor		
quantile-quantile	.457	.264	.216	.185
best fit	.384	.396	.311	.252
		City monitor		
quantile-quantile	.370	.232	.199	.170
best fit	.458	.441	.317	.243
		Sampling strategy of 1 day in 12		
		Watt Street monitor		
quantile-quantile	.276	.250	.205	.198
best fit	.510	.430	.310	.249
		Mounter Street monitor		
quantile-quantile	.485	.276	.202	.196
best fit	.428	.378	.286	.243
		City monitor		
quantile-quantile	.420	.249	.205	.163
best fit	.499	.431	.301	.233

Table 8.4: Average relative bias for the empirical quantile-quantile and 'best fit' models where the Watt Street, Mounter Street and City monitors are the complete data sets for 3 monitoring strategies, 1 day in 6, 1 day in 8 and 1 day in 12.

Model	x_{max}	x_{max-1}	x_{max-2}	x_{98}
Sampling strategy of 1 day in 6				
Watt Street monitor				
quantile-quantile	.078	.033	.056	.045
best fit	-.077	-.068	-.047	-.022
Mounter Street monitor				
quantile-quantile	-.211	-.043	.012	.047
best fit	-.008	-.036	-.025	-.012
City monitor				
quantile-quantile	-.064	-.019	-.029	-.023
best fit	-.096	-.097	-.057	-.029
Sampling strategy of 1 day in 8				
Watt Street monitor				
quantile-quantile	.079	.027	.050	.041
best fit	-.152	-.135	-.096	-.055
Mounter Street monitor				
quantile-quantile	-.229	-.057	-.004	.035
best fit	-.059	-.084	-.063	-.036
City monitor				
quantile-quantile	-.039	.007	-.001	.003
best fit	-.162	-.158	-.107	-.066
Sampling strategy of 1 day in 12				
Watt Street monitor				
quantile-quantile	.049	.003	.030	.020
best fit	-.071	-.060	-.035	-.011
Mounter Street monitor				
quantile-quantile	-.270	-.091	-.030	.010
best fit	.001	-.026	-.015	-.004
City monitor				
quantile-quantile	-.090	-.034	-.039	-.031
best fit	-.106	-.104	-.061	-.032

increases as sample size decreases indicating it is a more robust model. Clearly it is a matter of subjective judgement to decide when the intermittent sample size has become too small. For instance one may decide on various values of rmse above which the fit is not acceptable, for example .3, .4 or .5. It is probably best however, to make such a decision based on a consideration of both the graphs and the rmse factor, with the relative bias value being given less significance.

In summary then the quantile-quantile model would seem to be a very useful tool in air pollution monitoring. It does not require the assumption or identification of a statistical distribution for any of the data sets, it is simple to use as it requires simple mathematical techniques, and it would seem to be robust over a variety of sample sizes. In practical terms it requires a base data set from a site in the same airshed as intermittent monitoring site and its only theoretical assumption is that the distributional form, which need not be specified, is the same for both the restricted and complete data sets, an assumption which may be simply tested. It would also appear that the land use zones for the base and restricted monitoring sites should not differ too greatly.

8.7 Conclusions

Three statistical models have been considered for use in estimating upper percentiles of a restricted data set. The first is an empirical model linking the quantiles of this data set to that of a more complete data set at another site in the same airshed. The only assumption required is that the distributions of both data sets are the same but the parametric form need not be specified. The second model assumes a distributional form for the restricted data set, estimates the parameters of the distribution and thereby estimates the upper percentiles. The third approach is the same as the second except that the distributional form is first identified using a goodness-of-fit test.

It has been found that the first model (the quantile-quantile model) and the third (the best fit model) yield the best results when the three highest values and the 98-percentile values of 45 years of 24-h acid gas data are examined. In general, the quantile-quantile model yields better results, although it is possible that the model results are affected when the restricted and base data set sites correspond to very different land use zones. Four types of intermittent monitoring were considered: 1 day in 4, 1 in 6, 1 in 8, and 1 in 12. As expected the results worsened as the data sets became smaller, although the performance of the quantile-quantile model improved in comparison to the best fit model in general. These results and their simplicity indicate that the quantile-quantile model would appear to be a very useful tool in air quality management.

CHAPTER 9
SUMMARY AND CONCLUSIONS

In this thesis mathematical models have been developed for the purpose of air quality management. To reiterate, air quality management seeks to regulate the amount and location of pollutant emissions in such a manner as to satisfy some clearly defined set of ambient air quality standards ~~criteria~~. To achieve this goal what is required is a set of air quality standards ~~criteria~~, models to relate air quality to emissions, and monitoring data to determine the state of ambient air quality. In this thesis models have been constructed to predict the entire distribution of pollutant concentration. Prediction of the upper percentiles has been emphasised as it is these high pollutant concentrations that are most often referred to in air quality standards. Models have been developed both to relate source emissions of pollutants to receptor concentrations, and to aid in the development of monitoring programs which assess compliance with air quality standards ~~criteria~~ based upon complete and restricted data sets.

The models developed incorporate only as much complexity as is compatible with the intensiveness of the available data and the objectives of the modelling exercise. Accordingly, models have been formulated with as few parameters as necessary. The hybrid models developed in this thesis link two major approaches to air quality modelling in such a manner that the strengths of each approach are enhanced. The first component, the deterministic model, estimates the pollutant concentration based upon the causal variables such as emissions and meteorology. These models have been shown in numerous studies to perform poorly when estimating the upper percentiles of the distribution of pollutant concentration. However, they do produce reasonable estimates of the pollutant concentration about the median concentration. Using these data, the parameters of a previously identified distributional model may be estimated. The entire distribution of pollutant concentration can then be inferred.

It should be noted that hybrid modelling assumes that one distributional model is valid for the entire percentile range. While this assumption has not been verified theoretically it has been subjected to extensive empirical verification based upon the analysis of 55 acid gas data sets in Newcastle, 376 data sets recorded in Melbourne, for the carbon monoxide data set recorded near a line source and for sulphur dioxide observations recorded about an elevated point source.

In order to develop models describing the distribution of air quality data, the problem of estimation of the parameters of the appropriate distributional form for air quality data was examined in Chapter 3. In general, it was found that the method of maximum likelihood provided the best estimates of the upper percentiles, although this was achieved at the expense of greater computational demands. Other methods, such as the method of Menon (1963) for the Weibull distribution parameters, were identified as potentially useful where maximum likelihood methods could not be used. Importantly, it was demonstrated that approximate confidence intervals representing the minimum level of uncertainty could be easily constructed with the method of maximum likelihood.

Using the parameter estimation techniques described in Chapter 3, the problem of model identification was investigated in Chapter 4. Model identification was performed by comparing the maximum values of the respective log likelihood functions. Using simulation studies it was found, however, that the one-parameter exponential model could not be selected even where the data were distributed as exponential. The preferred procedure was in this case to use the Kolmogorov test statistics to verify that the exponential model was not applicable and then select between the remaining models using the likelihood function. The Kolmogorov test also allowed the hypothesis of particular distributional form describing the air quality data to be verified at the 95% confidence level.

Using the model identification procedure developed in Chapter 4, a data set consisting of observations of six pollutants recorded over several years at a number of sites within a large urban area was examined. It was found that no one distributional model was appropriate for all pollutants. These results, and those from the analysis of pollutant data sets in Chapters 5, 6 and 7, indicate the need to identify the distributional model applied to air quality data rather than selecting a model a priori. The importance of a model identification procedure for assessing compliance with air quality standards ~~criteria~~ based upon the analysis of incomplete data sets was also demonstrated in Chapter 8. Significant improvements in both the bias and variance of estimates of the upper percentiles were obtained compared with the a priori application of a single model.

With the techniques examined in Chapters 3 and 4, hybrid models were developed for three emission source regimes of importance in air quality management namely, area, line and point sources. The models were developed for inert or relatively inert pollutants and are applicable to large urban areas, roadway line sources and elevated point sources.

In Chapter 5 for acid gas levels in Newcastle, Australia, a simple deterministic model was linked with a gamma distributional model to provide a hybrid model predicting the entire distribution of pollutant concentration. It was demonstrated that a simple hybrid model could predict the upper percentiles with similar accuracy to that of the more complex model predictions of mean concentrations. The hybrid model included a procedure for generating approximate confidence intervals and these were found to provide reasonable bounds upon model uncertainty. Clearly this information will be useful in any management decision based upon model predictions.

A model predicting air pollutant concentrations dispersed from roadway line sources will be of interest in the assessment of motor vehicle emission policies, and the impacts associated with proposed developments and existing motor vehicle usage. In Chapter 6 a hybrid model was constructed which employed the Gaussian plume model and the Weibull distribution to predict the entire distribution of carbon monoxide concentrations. Again the hybrid model was found to predict the upper percentiles of the distribution of air quality observations well within the accuracy of a factor of 2 normally expected of air quality models. A model validation run confirmed the predictive ability of this model. The hybrid model also provided approximate confidence bounds for model estimates of the maximum concentration.

In Chapter 7 the problem of dispersion of pollutants from elevated point sources was considered. A point source Gaussian plume model incorporating the most recent refinements, where data were available, was employed to predict the distribution of pollutant concentration at a range of averaging times. Unlike the hybrid models developed for area and line sources it was necessary to calibrate the model output to a range of percentiles. This form of model calibration was required at all averaging times. The calibrated model was applied to three data sets in addition to the model calibration which confirmed the ability of the model to predict both the spatial and temporal variation of sulphur dioxide concentrations. Again the hybrid model was able to provide approximate 95% confidence intervals for model predictions which yielded reasonable estimates of model uncertainty.

Finally, in Chapter 8 the problem of assessing compliance with air quality standards based upon the analysis of restricted data sets was examined. This study employed the methods described in Chapters 3 and 4, of parameter estimation and model identification. Additionally it was demonstrated that an empirical quantile-quantile model could be used to assess compliance with air quality standards. However, to apply this model a second complete data set recorded within the same airshed is required. It was also found that, where possible, the data sets employed in the empirical quantile-quantile model should be drawn from similar land use categories. The empirical quantile-quantile model, while requiring an additional data set, was able to predict the upper percentiles of the distribution of air quality observations without the need to assume a particular distributional form. The assumption that both data sets are drawn from the same distributional form can be simply tested using a nonparametric test.

This thesis has shown that hybrid modelling techniques can provide reliable estimates of pollutant concentrations referred to by air quality standards. While this thesis has demonstrated that the hybrid modelling approach can be applied to several important areas of air quality management, it is not claimed that the studies reported here fully represent the range of possible applications of the hybrid modelling approach.

Considering the statistical component alone, an increase in the parameterization of statistical models may improve the performance of the simpler models when applied to complete or restricted data sets. A useful step would be to develop procedures for selecting between models of different parameterizations so that as few parameters as necessary are incorporated within the model. This should in turn lead to an improved assessment of compliance with air quality standards.

Finally, in order to determine more fully the advantages and the limitations of the hybrid modelling approach the models should be applied under the widest range of conditions possible. In particular, application of the hybrid models to the best data sets available would provide an indication of the ultimate limits of model accuracy.

REFERENCES

Anderson C.W. and Ray W.D. (1975) Improved maximum likelihood estimators for the gamma distribution. Communications in Statistics 4, 437-448.

Anscombe F.J. (1973) Graphs in statistical analysis. The American Statistician 27, 17-21.

Bach W. (1971) Variation of solar attenuation with height over an urbanized area. Journal of the Air Pollution Control Association 21, 621-628.

Bach W. (1978) The potential consequences of increasing CO_2 levels in the atmosphere. In Carbon dioxide, Climate and Society, J. Williams(ed.) IIASA/Pergamon, Oxford, 141-147.

Bain L.G. and Engelhardt M. (1980) Probability of correct selection of Weibull versus gamma based on likelihood ratio. Communications in Statistics A-9, 375-381.

Barry P.J. (1971) Use of argon-41 to study the dispersion of stack effluents. In Proceedings of the Symposium of Nuclear Techniques in Environmental Pollution, International Atomic Energy Agency, 241-253.

Beck M.B. (1981) Hard or soft environmental systems? Ecological Modelling 11, 233-251.

Benarie M.M. (1976) Urban air pollution modeling without computers. United States Environmental Protection Agency Publication No. EPA-600/4-76-055, Research Triangle Park, NC.

Benarie M.M. (1978) The simple box model revisited. Atmospheric Environment 12, 1929-1930.

Benarie M.M. (1980) The simple box model simplified. In Atmospheric Pollution 1980, Proceedings of the 14th International Colloquium, Paris, France, May 5-8, M.M. Benarie (ed.), Studies in Environmental Science, vol. 8, Elsevier, Amsterdam, 49-53.

Benarie M.M. (1982) Air pollution modelling operations and their limits. In Mathematical models for planning and controlling air quality, G. Fronza and P. Melli (eds) Pergamon, Oxford, 109-115.

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

Bencala K.E. and Seinfeld J.H. (1979) An air quality model performance assessment package. Atmospheric Environment 13, 1181-1185.

Benson P.E. (1982) Modifications to the Gaussian vertical dispersion parameter, σ_z, near roadways. Atmospheric Environment 16, 1399-1405.

Berger A., Melice J.L. and Demuth C.L. (1982) Statistical distributions of daily and high atmospheric SO_2 concentrations. Atmospheric Environment 16, 2863-2877.

Berman M. (1981) The maximum likelihood estimators of the gamma distribution are always positively biased. Communications in Statistics 10, 693-697.

Birnbaum Z.W. and Hall R.A. (1960) Small sample distributions for multisample statistics of the Smirnov type. The Annals of Mathematical Statistics 31, 710-720.

Bobee B. (1975) The log Pearson type 3 distribution and its application in hydrology. Water Resources Research 11, 681-689.

Brady G.I., Bower B.T. and Lakhami H.A. (1983) Estimates of the national benefits and costs of improving ambient air quality. Journal of Environmental Management 16, 191-210.

Briggs G.A. (1973) Diffusion estimation for small emissions. ~~In Environmental Research Laboratories, Air Resources Atmosphere Turbulence and Diffusion Laboratory 1973 annual report, USAEC Report ATDL-106, National Oceanic and Atmospheric Administration.~~ ATDL cont. file 79, ATDL, Oak Ridge.

Briggs G.A. (1975) Plume Predictions. In Lectures on air pollution and environmental impact analyses, American Meteorological Society, Boston, 59-111.

Businger J.A. (1982) Equations and Concepts. In Atmospheric Turbulence and Air Pollution Modelling, F.T.M. Nieunstadt and H. van Dop (eds), D. Reidel Publishing Co., Dordrecht, 1-36.

Campbell W.A. and Heath M.S. (1977) Air pollution legislation and regulations. In Air Pollution, vol. 5, A.C. Stern (ed..), Academic Press, New York, ~~3-40.~~ 355-379.

Carras J.N. and Williams D.J. (1981) Observations of near-field plume dispersion under extremely convective conditions. In Proceedings of the 7th International Clean Air Conference, K.A. Webb and A.J. Smith, (eds), Ann Arbor Science, 403-427.

Carras J.N. and Williams D.J. (1983) Observations of vertical plume dispersion in the convective boundary layer. In The 6th Symposium on Turbulence and Diffusion, American Meteorological Society, Boston, 249-252.

Cats G.J. and Holtslag A.A.M. (1980) Prediction of air pollution frequency distribution - Part I. The lognormal model. Atmospheric Environment 14, 255-258.

Chamberlain A.C. (1983) Effect of airborne lead on blood lead. Atmospheric Environment 17, 677-692.

Chambers J., Bridgman H.A. and Long G. (1982) Dispersion of a buoyant plume from tall stacks in the Middle Hunter Valley. Clean Air (Aust.) 16, 68-73.

Chambers J.M., Cleveland W.S., Kleiner B. and Tukey P.A. (1983) Graphical Methods for Data Analysis. Duxbury, Boston.

Chang T.Y. and Weinstock B. (1975) Generalized rollback modeling for urban air pollution control. Journal of the Air Pollution Control Association 25, 1033-1037.

Chock D.P. (1978) A simple line-source model for dispersion near roadways. Atmospheric Environment 12, 823-829.

Chock D.P. (1982a) Pollutant dispersion near roadways - experiments and modeling. Science of the Total Environment 25, 111-132.

Chock D.P. (1982b) Comment on "On the comparative assessment of the Performance of Air Quality Models". Journal of the Air Pollution Control Association 32, 282.

Chock D.P. (1984) Statistics of extreme values of a first-order Markov normal process: an exact result. Atmospheric Environment 18, 2461-2470.

Chock D.P. (1985a) A comparison of numerical methods for solving the advection equation - II. Atmospheric Environment 19, 571-586.

Chock D.P. (1985b) Personal Communication.

Choi S.C. and Wette R. (1969) Maximum likelihood estimation of the parameters of the gamma distribution and their bias. Technometrics 11, 683-690.

Cohen A.C. (1965) Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. Technometrics 7, 579-588.

Cohen A.C. and Norgaard N.J. (1977) Progressively censored sampling in the three-parameter gamma distribution. Technometrics 19, 333-340.

Conover N.J. (1980) Practical Nonparametric Statistics. 2nd Edition, John Wiley and Sons, New York.

Curran T.C. and Cox W.M. (1979) Data analysis procedures for the ozone NAAQS statistical format. Journal of the Air Pollution Control Association 29, 532-534.

Curran T.C. and Frank N.H. (1975) Assessing the validity of the lognormal model when predicting maximum air pollutant concentrations. Paper No. 75-51.3, presented at the 68th annual meeting of the Air Pollution Control Association, Boston.

D'Agostino R.B. (1971) An omnibus test of normality for moderate and large size samples. Biometrika 58, 341-348.

D'Agostino R.B. (1972) Small sample probability points for the D test of normality. Biometrika 59, 219-221.

Daly N.J. and Steele L.P. (1976) A predictive model for CO in Canberra. In Symposium on Air Pollution Diffusion Modeling, Australian Environment Council, Canberra, 264-275.

Davis W. and Metz D. (1978) A new technique for treatment of surface boundary conditions arising from particulate plume dispersion. Journal of Applied Meteorology 17, 1610-1618.

De Nevers N.H. and Morris J.R. (1975) Rollback modeling : basic and modified. Journal of the Air Pollution Control Association 25, 943-947.

De Nevers N.H., Neligan R.E. and Slater H.H. (1977) Air quality management, pollution control strategies, modeling and evaluation. In Air Pollution, vol.5, A.C. Stern (ed.), Academic Press, New York, 3-40.

Durbin J. (1975) Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. Biometrica *62*, 5-22.

Engeman R.M. and Keefe T.J. (1982) On generalized least squares estimation of the Weibull distribution. Communications in Statistics *11*, 2181-2193.

Environment Protection Authority of Victoria (1983) *Air monitoring results 1981*. Environment Protection Authority of Victoria, Melbourne, Australia.

Eschenroeder A. (1975) *An assessment of models for predicting air quality*. ERT Document ERTN-75-03, Environmental Research and Technology, Concord.

Ferris B.G. (1978) Health effects of exposure to low levels of regulated air pollutants: a critical review. Journal of the Air Pollution Control Association *28*, 482-497.

Fisher A.C. (1981) *Economic efficiency and air pollution control*. East-West Environment and Policy Institute Research Report No. 8, East-West Center, Honolulu.

Fox D.G. (1981) Judging air quality model performance. Bulletin of the American Meteorological Society *62*, 599-609.

Francey R.J. (ed.) (1984) *Baseline Atmospheric Program (Australia) 1981-82*. Department of Science and Technology, Canberra.

Georgopoulos P.G. and Seinfeld J.H. (1982) Statistical distributions of air pollutant concentrations. Environmental Science and Technology *16*, 401-416A.

Geraghty D. and Ricci P.F. (1984) Mathematical models for atmospheric pollutants. In *Principles of Health Risk Assessment*, P.F. Ricci (ed.), Prentice-Hall, Englewood Cliffs, New Jersey, 67-115.

Gibbons J.D. (1971) Nonparametric Statistical Inference. McGraw-Hill, New York.

Gifford F.A. (1974) The form of the frequency distribution of air pollution concentrations. In Proceedings of the Symposium on statistical aspects of Air Quality Data, United States Environmental Protection Agency Publication No. EPA-650/4-74-038, Research Triangle Park, NC.

Gifford F.A. and Hanna S.R. (1971) Urban Air Pollution Modeling. In Proceedings of the Second International Clean Air Congress, H.M. Englund and W.T. Beery (eds), Academic Press, New York, 1146-1151.

Gifford F.A. and Hanna S.R. (1973) Modelling urban air pollution. Atmospheric Environment 7, 131-136.

Gilchrist W. (1984) Statistical Modelling. John Wiley and Sons, Chichester.

Gilpin A. (1978) Air Pollution. 2nd edition, University of Queensland Press, St Lucia.

Giugliano M. (1985) The development and application of empirical models for the high SO_2 concentration in urban areas. The Science of the Total Environment 44, 89-96.

Green N.J. and Bullin J.A. (1982) Evaluation of roadway pollutant dispersion models using mass flux profiles and mass conservation. Environmental Science and Technology 16, 202-206.

Gualdi R. and Tebaldi S. (1982) Real-time control of air pollution: the case of Milan. In Mathematical Models for Planning and Controlling Air Quality, G. Fronza and P. Melli (eds), Pergamon, Oxford, 233-246.

Hanna S.R. (1971) A simple method of calculating dispersion from urban sources. Journal of the Air Pollution Control Association *21*, 774-777.

Hanna S.R. (1978) Diurnal variation of the stability factor in the simple ATDL urban dispersion model. Journal of the Air Pollution Control Association *28*, 147-150.

Hanna S.R. (1982a) *Review of Atmospheric Diffusion Models for Regulatory Applications*. World Meteorological Organization Technical Note No. 177, WMO-No. 581, Secretariat of the World Meteorological Organization, Geneva.

Hanna S.R. (1982b) Application in air pollution modelling. In *Atmospheric Turbulence and Air Pollution Modelling*, F.T.M. Nieuwstadt and H. van Dop (eds), D. Reidel Publishing Co., Dordrecht, 275-310.

Hanna S.R. (1982c) Natural variability of observed hourly SO_2 and CO concentrations in St. Louis. Atmospheric Environment *16*, 1435-1440.

Hanna S.R. and Gifford F.A. (1977) *Application of the ATDL simple urban dispersion model to Frankfurt, West Germany*. Atmospheric Turbulence and Diffusion Laboratory Publication No. 77/17, ATDL, Oak Ridge.

Hanna S.R., Egan B.A., Vaudo C.J. and Curreri A.J. (1984) A complex terrain dispersion model for regulatory applications at the Westvaco Luke Mill. Atmospheric Environment *18*, 685-699.

Harter H.L. and Moore A.H. (1965) Maximum liklihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples. Technometrics *7*, 639-643.

Hayes S.R. (1979) *Performance measures and standards for air quality simulation models*. United States Environmental Protection Agency Publication No. EF78-93R2, Research Triangle Park.

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

Holland D.M. and Fitz-Simons T. (1982) Fitting statistical distributions to air quality data by the maximum likelihood method. Atmospheric Environment 16, 1071-1076.

Hornberger G.M. and Spear R.C. (1980) Eutrophication in Peel inlet-I. The problem defining behaviour and a mathematical model for the phosphorus scenario. Water Research 14, 29-42.

Horowitz J. (1980) Extreme values from a nonstationary stochastic process: An application to air quality analysis. Technometrics 22, 469-478.

Horowitz J. and Barak~~ee~~at S. (1979) Statistical analysis of the maximum concentration of an air pollutant: Effects of autocorrelation and non-stationarity. Atmospheric Environment 13, 811-818.

Irwin J.S. (1979) Estimating plume dispersion - a recommended generalized scheme. In Preprints, Fourth Symposium on Turbulence, Diffusion and Air Pollution, American Meteorological Society, 62-69.

Jakeman A.J. and Simpson R.W. (1985) Assessment of air quality impacts from an elevated point source. Journal of Environmental Management 20, 63-72.

Jakeman A.J. and Young P.C. (1979) Refined instrumental variable methods of recursive time-series analysis, Part II: Multi-variable systems. International Journal of Control 29, 621-644.

Jakeman A.J. and Young P.C. (1981) On the decoupling of system and noise model parameter estimation in time-series analysis. International Journal of Control 34, 423-431.

Jakeman A.J. and Young P.C. (1983) Advanced methods of recursive time-series analysis. International Journal of Control *37*, 1291-1310.

Jakeman A.J., Greenaway M.A. and Jennings J.N. (1984) Time-series models for the prediction of stream flow in a karst drainage system. Journal of Hydrology (N.Z.) *23*, 21-33.

Jakeman A.J., Steele L.P. and Young P.C. (1980) Instrumental variable algorithms for multiple input systems described by mulitple transfer functions. IEEE Transactions on Systems, Man and Cybernetics *10*, 593-602.

Johnson G.T. (1980) *Sydney air quality modeling report.* Tech. Report No.80-004 School of Mathematics and Physics, Macquarie University, Sydney, Australia.

Johnson N.L. and Kotz S. (1970) *Continuous Univariate Distributions.* Houghton Mifflin, Boston.

Johnson W.B., Sklarew R.C. and Turner D.B. (1976) Urban air quality simulation modeling. In *Air Pollution,* A.C. Stern (ed), 3rd edition, vol. 1, Academic Press, New York, 503-562.

Jordan C. (1960) *Calculus of Finite Differences.* Chelsea Publishing Co., New York.

Kalpasanov Y. and Kurchatova G. (1976) A study of the statistical distribution of chemical pollutants in air. Journal of the Air Pollution Control Association *26*, 981-985.

Kappenman R.G. (1982) On a method for selecting a distributional model. Communication in Statistics *A-11*, 663-672.

Karplus W.J. (1976) The future of mathematical models of water resources systems. In *System Simulation in Water Resources,* G.C. Vansteenkiste (ed.), North-Holland, Amsterdam, 11-18.

Kendall M.G. and Stuart A. (1973) The Advanced Theory of Statistics. 3rd edition, Griffin, London.

Kent J.H. and Mudford N.R. (1978) Sydney driving patterns and automotive emission modeling. Charles Knolling Research Laboratory Tech Note ER-26, University of Sydney, Sydney, Australia.

Kim T.J. and Hoskote N.G. (1983) Estimating mobile source pollutant emission: methodological comparison and planning implications. Environmental Monitoring and Assessment 3, 1-12.

Knox J.B. and Lange R. (1974) Surface air pollutant concentration frequency distributions: Implications for urban modeling. Journal of the Air Pollution Control Association 24, 48-53.

Lamb R.G. and Seinfeld J.H. (1973) Mathematical modeling of urban air pollution: General theory. Environmental Science and Technology 7, 253-261.

Larsen R.I. (1969) A new mathematical model of air pollutant concentration averaging time and frequency. Journal of the Air Pollution Control Association 19, 24-30.

Larsen R.I. (1971) A mathematical model for relating air quality measurements to air quality standards. United States Environmental Protection Agency Publication No. AP-89, Research Triangle Park, NC.

Larsen R.I. (1973) An air quality data analysis system for interrelating effects, standards and needed source reductions. Journal of the Air Pollution Control Association 23, 933-940.

Larsen R.I. (1974) An air quality data analysis system for interrelating effects, standards and needed source reductions - II. Journal of the Air Pollution Control Association 24, 551-558.

Larsen R.I. and Heck W.W. (1976) An air quality data analysis system for interrelating effects, standards, and needed source reductions: Part 3. Vegetation injury. Journal of the Air Pollution Control Association 26, 325-333.

Larsen R.I. and Heck W.W. (1984) An air quality data analysis system for interrelating effects, standards, and needed source reductions: Part 8. An effective mean O_3 crop reduction mathematical model. Journal of the Air Pollution Control Association 34, 1023-1034.

Lawlor L. (1982) An assessment of motor vehicle air pollution in Australia 1975-1980. Bureau of Transport Economics, Canberra.

Lemon G.H. (1975) Maximum likelihood estimation for the three parameter Weibull distribution based on censored samples. Technometrics 17, 247-254.

Lilliefors H.W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399-402.

Liu M. and Moore G.E. (1984) On the evaluation of predictions from a Gaussian plume model. Journal of the Air Pollution Control Association 34, 1044-1050.

Looney S.W. and Gulledge T.R. (1985) Use of the correlation coefficient with normal probability plots. The American Statistician 39, 75-79.

Lynn D.A. (1974) Fitting curves to urban suspended particulate data. In Proceedings of the Symposium on Statistical Aspects of Air Quality Data, United States Environmental Protection Agency Publication No. EPA-650/4-74-038, Research Triangle Park, NC.

Maccarrone S. (1985) Air quality in the vicinity of roads. In Proceedings of the 13th Conference of the Australian Road Research Board (to appear).

Maddukuri C.S. (1982) A numerical model of diffusion of Carbon Monoxide near highways. Journal of the Air Pollution Control Association 32, 834-836.

Mage D.T. (1980) Frequency distributions of hourly wind speed measurements. Atmospheric Environment 14, 367-374.

Mage D.T. and Ott W.R. (1978) Refinements of the lognormal probability model for analysis of aerometric data. Journal of the Air Pollution Control Association 28, 796-798.

Mage D.T. and Ott W.R. (1984) An evaluation of the method of fractiles, moments and maximum likelihood for estimating parameters when sampling air quality data from a stationary lognormal distribution. Atmospheric Environment 18, 163-171.

Mainwaring S.J. and Thorpe J.L. (1983) Predictions of line-source emissions of particulate lead. Clean Air (Aust.) 17, 67-69.

Menon M.V. (1963) Estimation of the shape and scale parameters in the Weibull distribution. Technometrics 5, 175-182.

Naylor T.H., Balintfy J.L., Burdick D.S. and Chu K. (1966) Computer Simulation Techniques. John Wiley and Sons, New York.

Newill V.A. (1977) Air Quality standards. In Air Pollution, Vol. 5. A.C. Stern (ed.), Academic Press, New York, 445-504.

Nieuwstadt F.T.M. (1980) Prediction of air pollution frequency distribution - Part II. The Gaussian plume model. Atmospheric Environment 14, 259-265.

Nozdryn-Plotnicki M.J. and Watt W.E. (1979) Assessment of fitting techniques for the log Pearson type 3 distribution using Monte Carlo simulation. Water Resources Research 15, 714-718.

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.

Ott W.R. and Mage D.T. (1981) Measuring air quality levels inexpensively at multiple locations by random sampling. Journal of the Air Pollution Control Association 31, 365-369.

Ott W.R., Mage D.T. and Randecker V.W. (1979) Testing the Validity of the Lognormal Probability Model: Computer Analysis of Carbon Monoxide Data from U.S. Cities. United States EPA Office of Research and Development, Publication No. EPA-600/4-79-040, Washington DC.

Pack D.H. (1982) Precipitation chemistry probability - the shape of things to come. Atmospheric Environment 16, 1145-1157.

Pasquill F. and Smith F.B. (1983) Atmospheric Diffusion. 3rd edition, Ellis Horwood, Chichester.

Pearson E.S., D'Agostino R.B. and Bowman K.O. (1977) Tests for departure from normality: Comparison of powers. Biometrika 64, 231-246.

Peterson T.W. and Moyers J.L. (1980) Emission limits for variable sources by use of multipoint rollback. Atmospheric Environment 14, 1439-1444.

Pierce T.E. (1984) An evaluation of alternative Gaussian plume dispersion modeling techniques in estimating short-term sulfur dioxide concentrations. United States Environmental Protection Agency Publication No. EPA-600/53-84-079, Research Triangle Park, NC.

Pollack R.I. (1975) Studies of pollutant concentration frequency distributions. United States Environmental Protection Agency, Publication No. 650/4-75-004, Research Triangle Park, NC.

Rao S.T. and Visalli J.R. (1981) On the comparative Assessment of the performance of air quality models. Journal of the Air Pollution Control Association 31, 851-860.

Roberts E.M. (1979a) Review of statistics of extreme values with applications to air quality data: Part I. Review. Journal of the Air Pollution Control Association 29, 632-637.

Roberts E.M. (1979b) Review of statistics of extreme values with applications to air quality data: Part II. Applications. Journal of the Air Pollution Control Association 29, 733-740.

Rodden J.B., Green N.J., Messina A.D. and Bullin J.A. (1982) Comparison of roadway pollutant dispersion models using the Texas data. Journal of the Air Pollution Control Association 32, 1226-1228.

Rosher V., Pitt D., Banbury E. and Papro V. (1984) Kalgoorlie air quality investigation : Analysis of data July 1982-March 1984. Environmental Note No. 150, Department of Conservation and Environment, Perth, Western Australia.

Scriven R.A. (1971) Use of argon-41 to study the dispersion of stack effluents. In Proceedings of the Symposium on Nuclear Techniques in Environmental Pollution. International Atomic Energy Agency, 254-255.

Seinfeld J.H. (1975) Air Pollution: Physical and Chemical Fundamentals. McGraw Hill, New York.

Shapiro S.S. and Francia R.S. (1972) Approximate analysis of variance test for normality. Journal of the American Statistical Association 67, 215-216.

Shapiro S.S. and Gross A.J. (1981) Statistical Modeling Techniques. Marcel Dekker, New York.

Shapiro S.S. and Wilk M.B. (1965) An analysis of variance test for normality (complete samples). Biometrika 52, 591-611.

Shapiro S.S., Wilk M.B. and Chen H.J. (1968) A comparative study of various tests for normality. Journal of the American Statistical Association 63, 1343-1372.

Shenton L.R. and Bowman K.O. (1972) Further remarks on maximum likelihood estimators for the gamma distribution. Technometrics 14, 725-733.

Siddiqi T.A. and Chong-Xian Z. (1984) Ambient air quality standards in China. Environmental Management 8, 473-479.

Simpson R.W. (1984) Predicting Frequency Distributions for ozone, NO_2 and TSP from restricted data sets. Atmospheric Environment 18, 353-360.

Simpson R.W. and Hanna S.R. (1981) A review of deterministic urban air quality models for inert gases. U.S. National Office of Air Assessment, Technical Memo ERL ARL-106, Maryland.

Simpson R.W. and Jakeman A.J. (1984) A model for estimating the effects of fluctuations in long-term meteorology on observed maximum acid gas levels. Atmospheric Environment 18, 1633-1640.

Simpson R.W. and Jakeman A.J. (1985) Forecasting worst case pollution scenarios for acid gas and suspended particulates due to urban industrial development. Environmental Pollution (Series B) 9, 137-149.

Simpson R.W., Butt J. and Jakeman A.J. (1984) An averaging time model of SO_2 frequency distributions from a single point source. Atmospheric Environment 18, 1115-1123.

Simpson R.W., Daly N.J. and Jakeman A.J. (1983) The prediction of maximum air pollution concentrations for inert gases using Larsen's model and the ATDL model. Atmospheric Environment 17, 2497-2503.

Simpson R.W., Jakeman A.J. and Daly N.J. (1985) The relationship between the ATDL model and the statistical distributions of wind speed and pollution data. Atmospheric Environment 19, 75-82.

Singpurwalla N.D. (1972) Extreme values from a lognormal law with applications to air pollution problems. Technometrics 14, 703-711.

Sinha S.K. and Kale B.K. (1980) Life Testing and Reliability Estimation. Wiley, New Delhi.

Sistla G., Samson P., Keenan M. and Rao S.T. (1979) A study of pollutant dispersion near highways. Atmospheric Environment 13, 669-685.

Stedinger J.R. (1980) Fitting log normal distributions to hydrologic data. Water Resources Research 16, 481-490.

Steele L.P. (1981) Recursive estimation in the identification of air pollution models. Ph.D. Thesis, Australian National University.

Steele L.P. and Jakeman A.J. (1980) Recursive methods of time series analysis for modeling atmospheric environments. Proceedings of the Simulation Society of Australia, Biennial Conference, Brisbane, August 27-29, 75-82.

Stern A. (1976) Air Pollution. vol. 3, Academic Press, New York.

Surman P.G., Simpson R.W. and Stokoe J. (1982) Application of Larsen's model to Brisbane, Australia. Atmospheric Environment 16, 2609-2614.

Sutton O.G. (1932) A theory of eddy diffusion in the atmosphere. Proceedings of the Royal Society A 135, 143-165.

Sutton O.G. (1934) Wind structure and evaporation in a turbulent atmosphere. Proceedings of the Royal Society A 146, 701.

Taylor G.I. (1915) Eddy motion in the atmosphere. Philosophical Transactions of the Royal Society A 215, 1.

Taylor G.I. (1921) On the theory of diffusion by continuous movements. Proceedings of the London Mathematical Society 20, 196-212.

Taylor G.I. (1927) Turbulence. Quarterly Journal of the Royal Meteorological Society 53, 201.

Thoman D.R., Bain L.J. and Antle C.E. (1969) Inferences on the parameters of the Weibull distribution. Technometrics 11, 445-460.

Tiao G.C. and Hillmer S.C. (1978) Statistical models for ambient concentrations of carbon monoxide, lead, and sulfate based on the LACS data. Environmental Science and Technology 12, 820-828.

Tong E.Y. and De Pietro S.A. (1977) Sampling frequencies for determining long-term average concentrations of atmospheric particulate sulfates. Journal of the Air Pollution Control Association 27, 1008-1011.

Trijonis J. (1978) Empirical Relationships between Atmospheric Nitrogen Dioxide and its Precursors. United States Environmental Protection Agency Publication No. 600/3-78-018, Research Triangle Park, NC.

Tsukatani T. and Shigemitsu K. (1980) Simplified Pearson distributions applied to air pollutant concentration. Atmospheric Environment 14, 245-253.

Turner D.B. (1979) Atmospheric dispersion modeling: A critical review. Journal of the Air Pollution Control Association 29, 502-519.

Turner D.B. and Irwin J.S. (1982) Extreme value statistics related to performance of a standard air quality simulation model using data at seven power plant sites. Atmospheric Environment 16, 1907-1914.

United States National Research Council (1976) Halocarbons: Effects on stratospheric ozone. National Academy of Sciences, Washington, DC.

United States National Research Council (1983a) Acid Deposition: Atmospheric Processes in Eastern Europe. National Academy of Sciences, Washington, DC.

United States National Research Council (1983b) Changing Climate. National Academy of Sciences, Washington, DC.

Vemuri V. (1978) Modeling of complex systems: An introduction. Academic Press, New York.

Venkatram A. (1983) Uncertainty in predictions from air quality models. Boundary-Layer Meteorology 27, 185-196.

Venkatram A. (1984) The uncertainty in estimating dispersion in the convective boundary layer. Atmospheric Environment 18, 307-310.

Venkatram A. and Pleim J. (1985) Analysis of observations relevant to long-range transport and deposition of pollutants. Atmospheric Environment 19, 659-667.

Watson I.D. (1983) Line-source model predictions of carbon monoxide concentrations in Sydney. Clean Air (Aust.) 17, 72-76.

Weibull W. (1951) A distribution function of wide applicability. Journal of Applied Mechanics 18, 293-297.

Wilk M.B. and Gnanadesikan R. (1968) Probability plotting methods for the analysis of data. Biometrika 55, 1-17.

World Health Organization (1972) Air Quality Criteria and Guide for Urban Pollutants. Technical Report No. 506, World Health Organization, Geneva.

Wyzga R.E. (1978) The effect of air pollution upon mortality: A consideration of distributed lag models. Journal of the American Statistical Association 73, 463-472.

Young P.C. (1974) Recursive approaches to time-series analysis. Institute of Mathematics and its Applications 10, 209-224.

Young P.C. (1976) Some observations on instrumental variable methods of recursive time-series analysis. International Journal of Control 23, 593-612.

Young P.C. (1978) A general theory of modeling for badly defined systems. In Modeling, identification and control in environmental systems, G.C. Vansteenkiste (ed.), North-Holland, Amsterdam.

Young P.C. (1982a) The validity and credibility of models for badly defined systems. In Uncertainty and forecasting of water quality, M.B. Beck and G. van Straten (eds), IIASA/Pergamon, Oxford.

Young P.C. (1982b) Systems methods in the evaluation of environmental pollution problems. In Pollution: causes, effects and control, R.M. Harrison (ed.), Royal Society of Chemistry,London.

Young P.C. and Jakeman A.J. (1979) Refined instrumental variable methods of recursive time-series analysis, Part I: Single input, single output systems. International Journal of Control 29, 1-30.

Young P.C. and Jakeman A.J. (1980) Refined instrumental variable methods of recursive time-series analysis, Part III: extensions. International Journal of Control 31, 741-764.

Zimmerman J.R. and Thompson R.S. (1975) User's guide for HIWAY, a highway air pollution model. United States Environmental Protection Agency, Publication No. EPA-650/4-74-008, Research Triangle Park, NC.