A SURVEY OF STATISTICAL TECHNIQUES

USED IN VALIDATION STUDIES OF AIR POLLUTION PREDICTION MODELS

by

Robert D. Bornstein

and

Steven F. Anderson

Department of Meteorology
San Jose State University
San Jose, CA 95192

9509691

TECHNICAL REPORT NO. 23

MARCH 1979

STUDY ON STATISTICS AND ENVIRONMENTAL

FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)

ROCKEFELLER FOUNDATION

SLOAN FOUNDATION

ENVIRONMENTAL PROTECTION AGENCY (EPA)

NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISCLAIMER

DISCLAIMER

TABLE OF CONTENTS

A Survey of Statistical Techniques Used in Validation Studies of Air Pollution Prediction Models

by

Robert D. Bornstein

and

Steven F. Anderson

Department of Meteorology
San Jose State University

Abstract

Statistical techniques used by meteorologists to validate predictions made by air pollution models are surveyed. Techniques are divided into the following three groups: graphical, tabular, and summary statistics. Some of the practical problems associated with verification are also discussed. Characteristics desired in any validation program are listed and a suggested combination of techniques which possesses many of these characteristics is presented.

List of Symbols

C	concentration
E	fractional error
E_p	fractional percent error
N	number of observations
r	linear correlation coefficient
r'	logarithmic correlation coefficient
R	relative root mean square error
S	standard error of the estimate
t	time
ϵ	normalized root mean square error
σ	root mean square value
σ_Δ	σ of observed minus predicted concentration
σ_E	σ of relative concentration
$\Delta(\ )$	difference operator
$\lvert(\ )\rvert$	absolute value operator
$(\overline{\ })$	averaged value
$(\ )_0$	observed value
$(\ )_p$	predicted value

I. Introduction

Air pollution is a problem that can only be solved by a coalition of scientists and nonscientists. For example, educating the public in order that they might let elected officials know they are willing to pay the price for cleaning the environment is, at least partially, a role of public-interest groups, while passing laws requiring reduced emissions is a political problem. In the scientific area, the actual reduction of emissions is an engineering problem, the transport and diffusion of pollutants in the atmosphere is a meteorological problem, while transformation of atmospheric pollutants is a chemical problem.

Possible contributions of statisticians in cleaning the atmospheric environment include use of statistical techniques to predict pollutant concentrations, design of optimum networks and optimum sampling procedures to collect concentration data, verification of pollution prediction models, and the correlation of adverse health effects with atmospheric pollutants.

For a variety of reasons little past communication has occurred between statisticians and air pollution modelers. Because of this poor communication, modelers have not had access to the latest analysis techniques developed by statisticians and have not had opportunities to influence development of the new analysis techniques that statisticians could produce if they were aware of the specific problems facing air pollution modelers.

This paper, therefore, surveys some techniques presently used by air pollution modelers so as to allow statisticians to estimate the approximate level of knowledge existing in the air pollution modeling community. Also discussed are practical problems associated with the verification of air pollution models. This information should be useful to statisticians hoping to develop verification procedures needed to overcome these problems.

II. Model Validation Techniques

A. Graphical

Graphical comparisons between observed and predicted concentrations can be carried out using values at a single site over an extended time period or values at various sites at a single time. An example of the former is the presentation of superimposed time series plots of observed and predicted concentrations. This was done by Johnson et al. (1970) to validate a carbon monoxide highway prediction model (Figure 1).

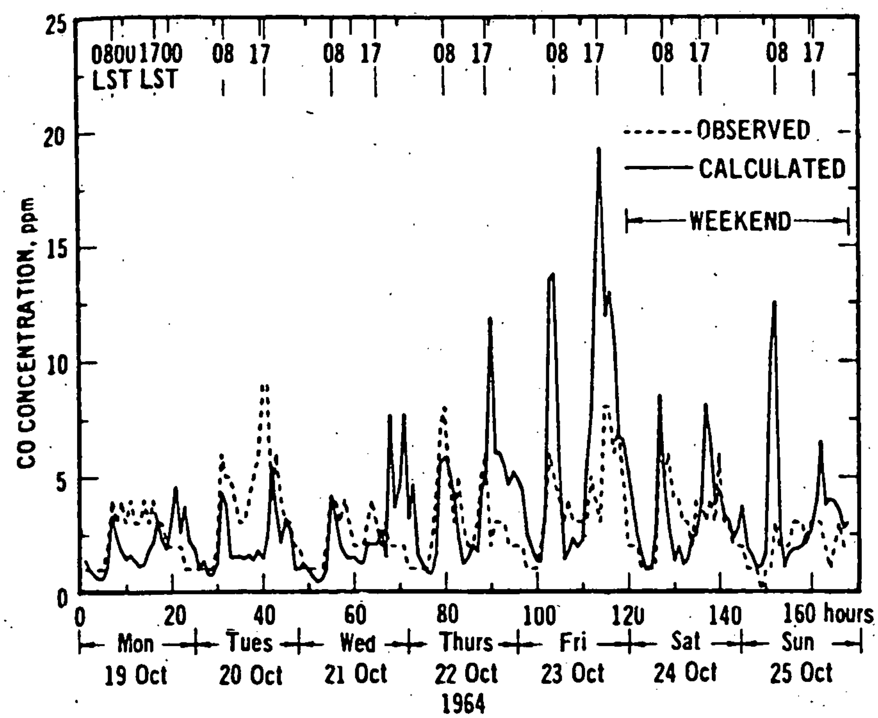


Figure 1. Hourly predicted and observed CO concentrations at St. Louis CAMP station for period 19-25 October, 1964. (From Johnson et al., 1970)

A similar approach was used by Roberts et al. (1970) for results from a Gaussian puff sulfur dioxide model (Figure 2). However, in this application observed concentration values C were smoothed (in order to increase "readability") by use of

$$(1) \qquad \overline{C}(t) = 0.25[C(t-\Delta t) + 2C(t) + C(t+\Delta t)] \; ,$$

where Δt was taken as one hour. Another variation of the time series plot technique used by Shir and Sheih (1973) utilized output from a numerical sulfur dioxide model and plotted observational data points (Figure 3). Use of a logarithmic concentration axis reduces the scatter between observed and simulated values.

Time series plots allow for qualitative evaluation of how well models reproduce: 1) the magnitude of extreme (maximum and minimum) values, 2) the time of occurrence of extreme values, 3) daytime (unstable) versus nighttime (stable) differences, and 4) weekday (high source strength) versus weekend (low source strength) differences. The method by itself, however, does not provide quantitative estimates of model performance in general, or performance under the various specialized conditions listed above.

The ability of a model to predict variations of surface pollutant concentration across a city can be demonstrated by constructing transections, as was done by Shieh et al. (1970) with output from a Gaussian sulfur dioxide model. As shown in Figure 4 this technique provides a good qualitative summary of how well a model simulates regions with alternating strong and weak (i.e., Central Park, which extends from CPW to 5th Avenue) area source emissions.

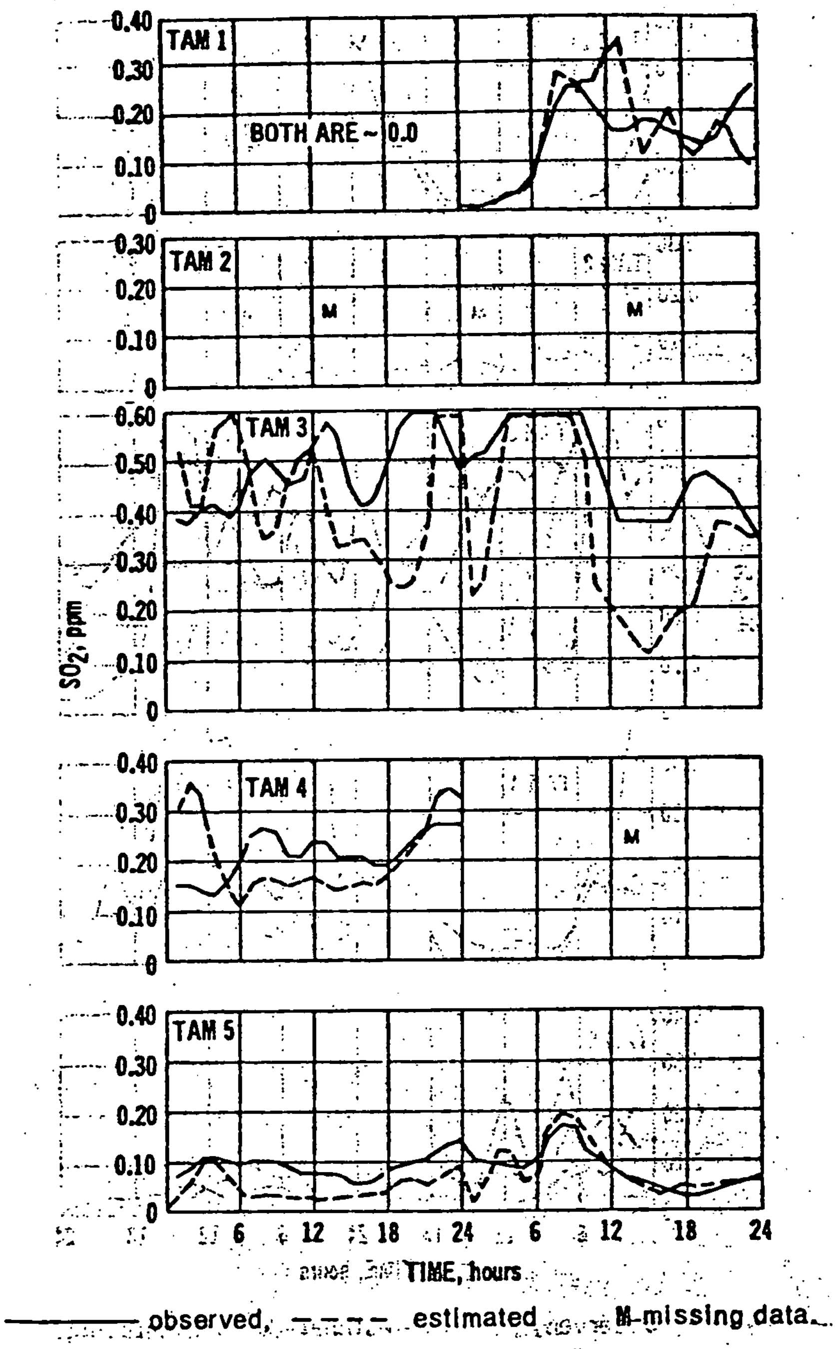


Figure 2. Sample of smoothed hourly values of SO_2 concentrations at TAM stations 1 to 5, January 3 and 4, 1967. (From Roberts et al., 1970)

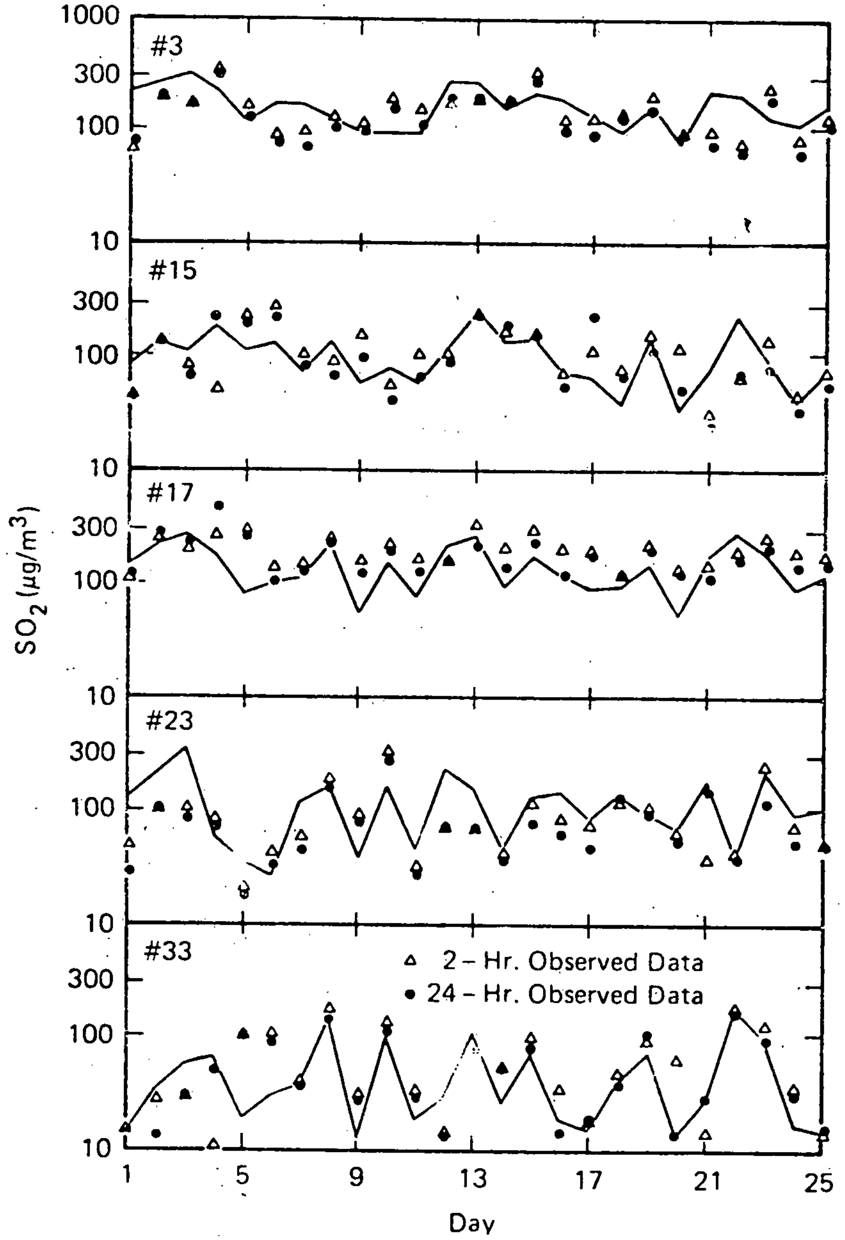


Figure 3. Comparison of observed vs. computed (solid line) 2-hour and 24-hour averaged variations of SO_2 concentrations for 25 days period. (From Shir and Shieh, 1973)

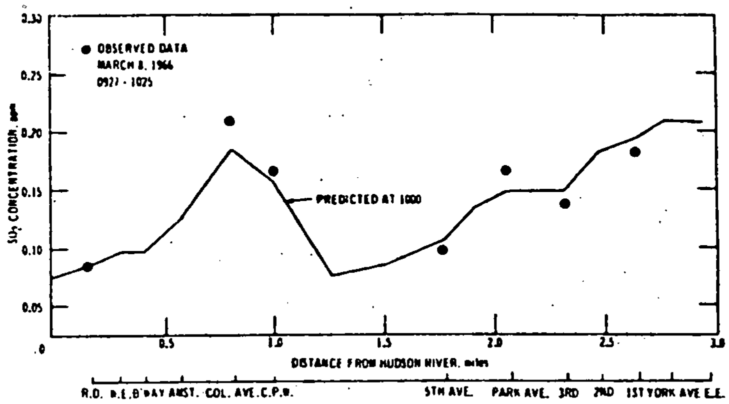


Figure 4. Observed and predicted variation of SO_2 concentration along the 79th Street crosstown transect at the indicated times. (From Shieh et al., 1970)

This technique is also frequently applied in studies of concentration profiles associated with plumes from large individual sources. Concurrent observed and simulated profiles are presented either in the direction along the plume axis, in the vertical at a given downwind distance, or in the lateral direction perpendicular to the plume axis at a given downwind distance. Examples of the first type for a surface-based source are shown in Figure 5, from Stephens and McCaldin (1971), and for an elevated source in Figure 6, from Slade (1968). Note the logarithmic concentration scale in the former figure again reduces the scatter of the observations about the calculated curves.

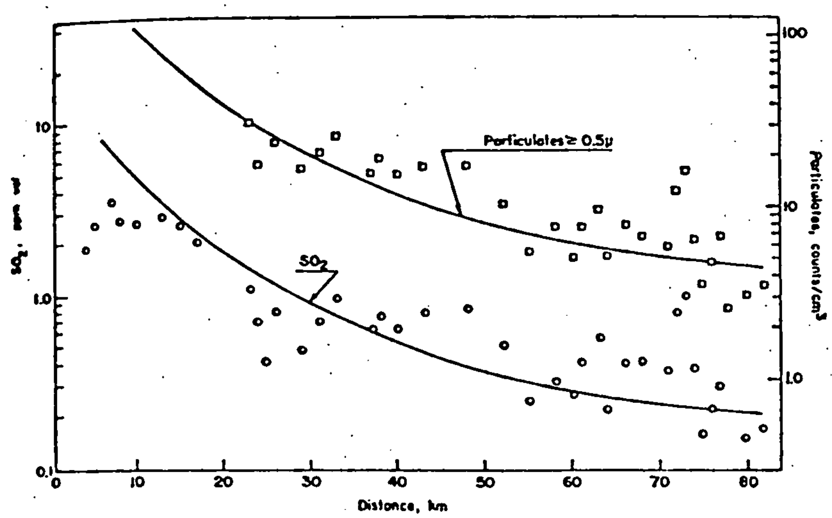


Figure 5. Calculated and experimental concentration profiles on Jan. 3, 1969. (Taken from Stephens and McCaldin, 1971)

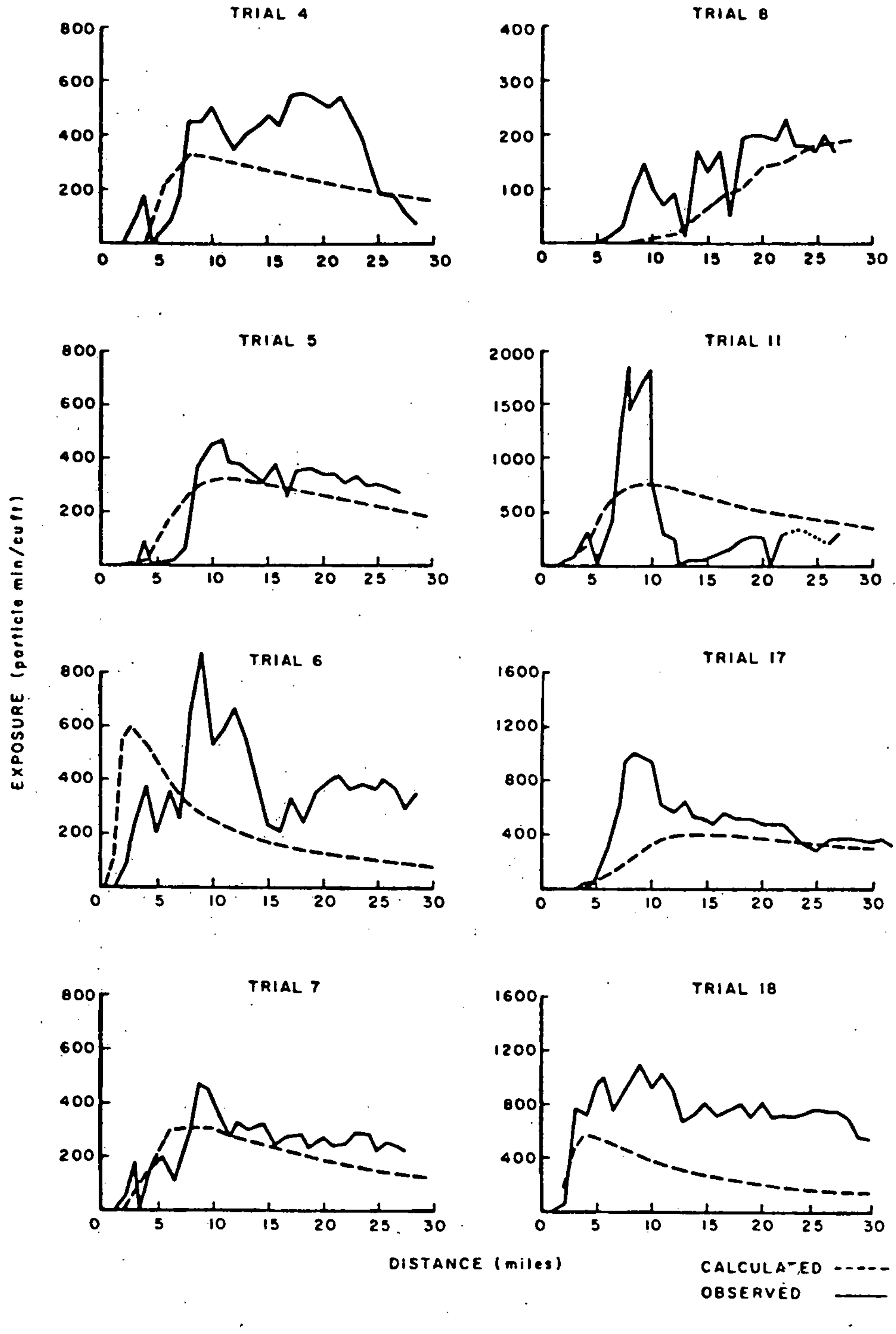


Figure 6. Comparison of calculated and observed exposures from an elevated line-source release. (Taken from Slade, 1968)

Transections can provide good qualitative estimates of how well a model reproduces the magnitude and location of extreme values under a variety of source strength and meteorological conditions. However, as with time series plots, transections by themselves do not provide quantitative estimates of model performance.

The ability of a model to simulate spatial distributions of pollutants in vertical or horizontal planes can be qualitatively shown by comparing simultaneous observed and predicted isopleth analyses for pollution studies involving multiple urban sources or single large point sources. Such comparisons were made by Dietzer (1976) using predicted surface concentrations in the Ruhr area obtained from a statistical model based on eigenvectors and associated time dependent eigencoefficients (Figure 7).

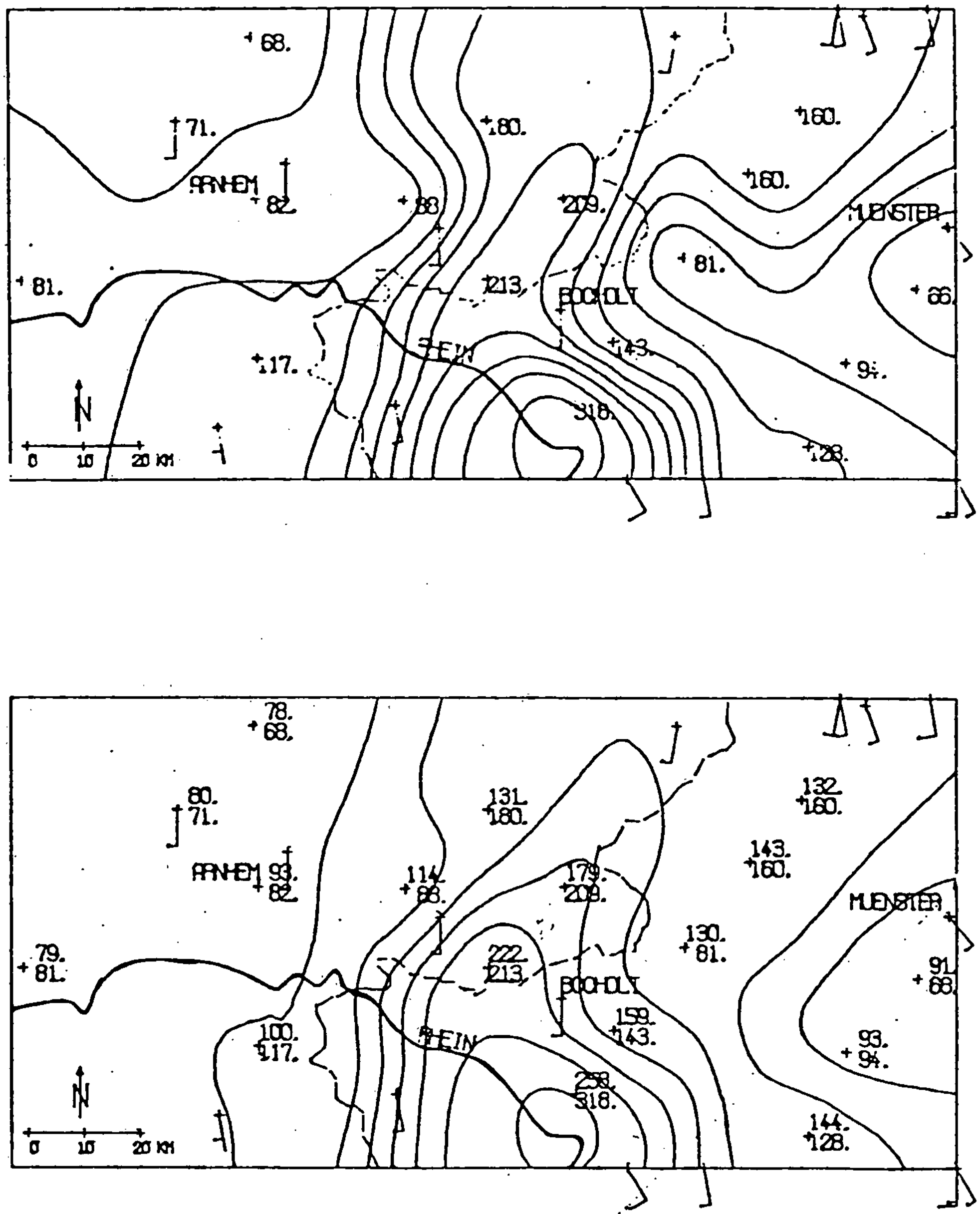


Figure 7. Observed (upper) and computed (lower) concentration fields.
(Taken from Dietzer, 1976)

A similar procedure can be used to investigate the ability of a model to simulate concentration fields in a vertical plane across a city, as was done by Shir and Sheih (1973). However, they lacked the vertical soundings necessary to construct observed concentration cross sections, and thus only presented simulated cross sections (Figure 8).

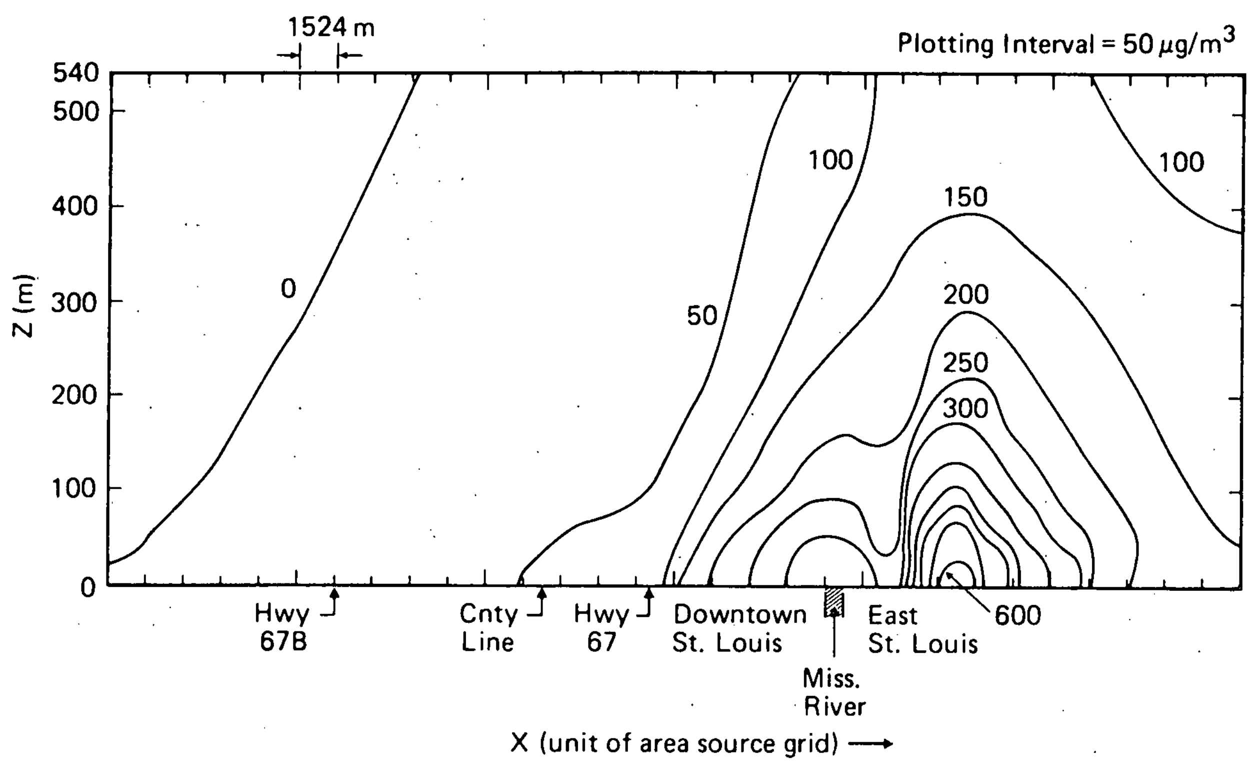
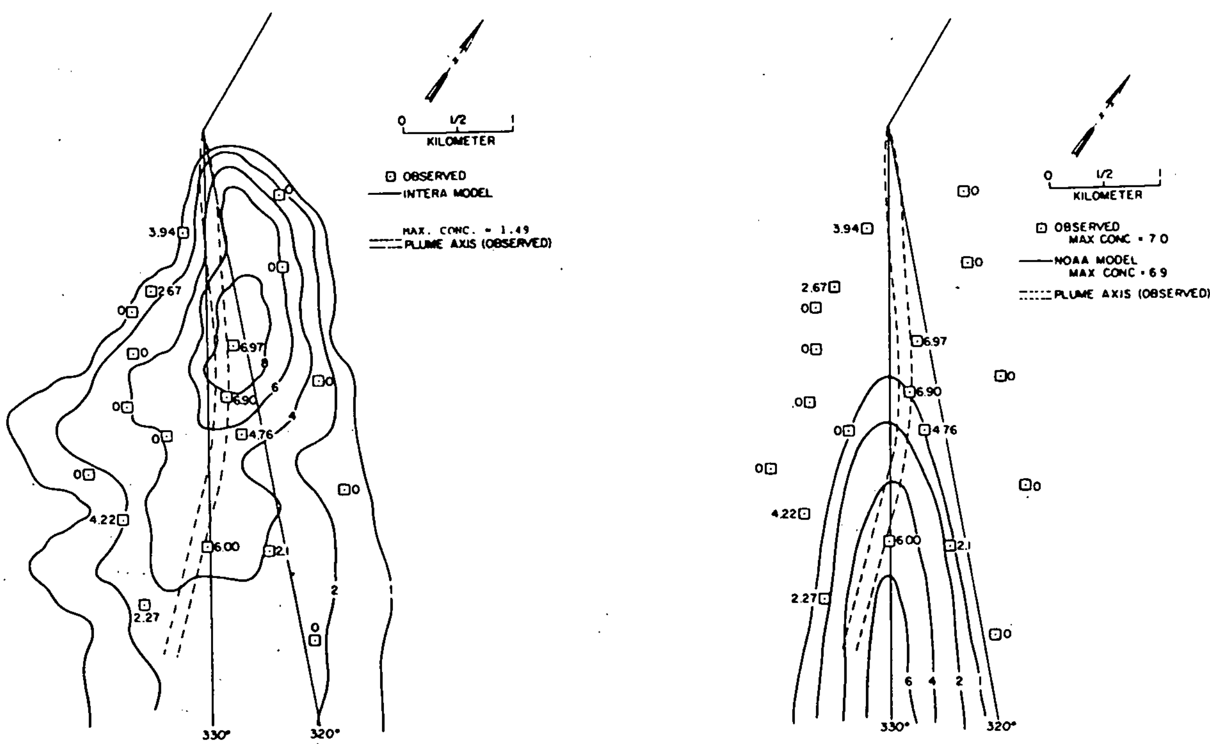


Figure 8. Simulation of hourly average SO_2 concentration field in X-Z plane. (Taken from Shir and Shieh, 1973)

Isopleths of predicted ground level concentrations resulting from an elevated point source were presented by Lantz et al. (1976) along with plotted observed concentration values (Figure 9). This allowed for a visual, qualitative estimate of model accuracy.

A possible extension of concurrent plotted predicted and observed concentration fields involves isoplething differences between observed and computed values. This allows for evaluation of model accuracy as a function of position relative to significant sources.



A. NUMERICAL MODEL RESULTS B. NOAA MODEL RESULTS

Figure 9. Comparison of ground-level concentration patterns at Garfield, Utah, with values in 10^{-7} gm m^{-3}. (Taken from Lantz et al., 1976)

Isopleth analysis evaluates the ability of a model to simulate basic spatial concentration patterns, even though predicted locations of maximum and minimum values may be displaced in space by, for example, inaccuracies in the input wind direction near a large elevated point source. For the purposes of studies of health effects and air quality management, these results may still be very useful, but comparisons between predicted and observed concentrations at a particular point would yield poor results.

This effect was quantified by Tesche et al. (1979) by use of a frequency distribution plot of the distance (in number of grid cells) from each observational site at which predicted concentrations first equaled observed values (Figure 10). The modal value can be seen to be only one grid interval.

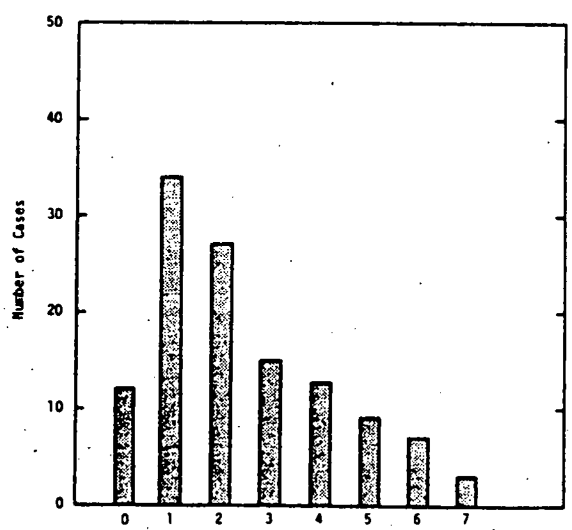


Figure 10. Number of grid cells for which model predictions bracket observed ozone concentrations for 18 stations during the period 1000 to 1700 PST. (Taken from Tesche et al., 1979)

The space scales associated with the grid sizes used in all air pollution prediction models introduce problems when observed and predicted concentration fields are compared. For example Shieh et al. (1970) demonstrated that a smaller area source emission grid gave a concentration field with greater detail than one using a larger area source emission grid (compare Figures 11 and 12). This illustrates the basic problem associated with comparisons of predicted volume averaged values to observed point values.

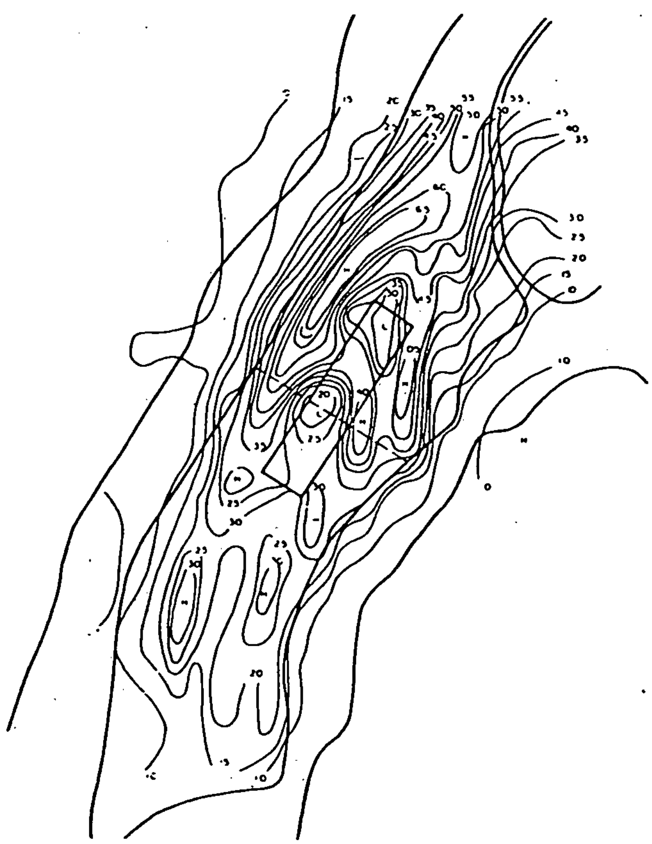


Figure 11. Predicted SO_2 concentration field (ppm), 1800 March 9, 1966, 0.2 x 0.2 mile computational grid. (From Shieh et al., 1970)

Figure 12. Predicted SO_2 concentration field (ppm), 1800 March 9, 1966, 1.0 mile x 1.0 mile computational grid. (From Shieh et al., 1970)

The ultimate aspect of this problem is the tremendous variation in concentration that can occur (Figure 13) at the intersection of two streets (i.e., within an urban canyon) due to complex microscale circulation patterns (Figure 14). Most observations of pollutant concentration are within the canyon (i.e., point observations at street level) but the output from air pollution prediction models are volume averages for "above rooftop levels," as the models must ignore the complex urban topography associated with individual buildings.

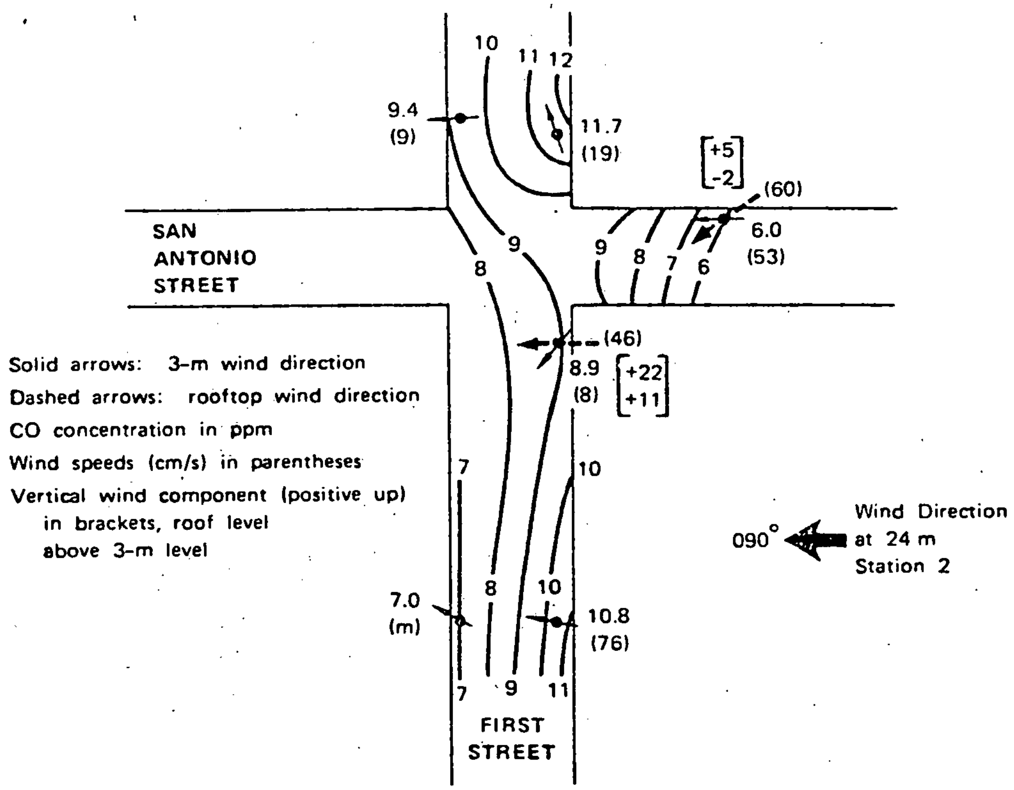


Figure 13. Concentration variation at the intersection of two street corners. (Taken from Johnson et al., 1971)

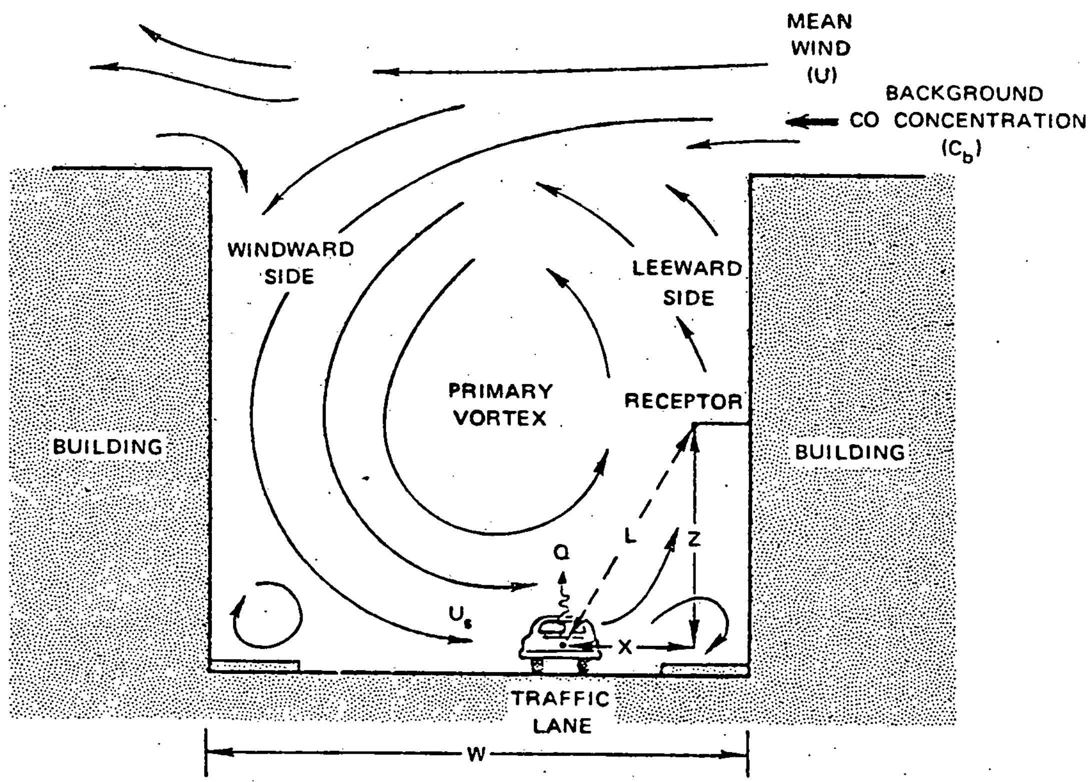


Figure 14. Schematic of cross-street circulation between buildings. (Taken from Johnson et al, 1971)

To overcome this problem, Johnson et al. (1970) pointed out the need to relate predicted rooftop concentrations to street level observed values using wind and traffic density data. For the purposes of constructing concurrent isopleth analyses of observed and computed concentration fields, a way should be developed to bring the observed canyon values "up to" roof level, while for the quantitative statistical comparisons discussed in Section C, a way should be developed to bring the predicted rooftop values "down" to the canyon street level.

This problem of scale is part of an even more general problem for which input from statisticians can be very useful, i.e., the problem of network design. Quantitative attempts have been made to develop criteria for placing observing stations within air sheds with respect to required number of stations and their optimum location. Statisticians might be able to provide input into developing such criteria and in developing criteria for optimum sampling frequencies.

A scatter diagram is another graphical technique for comparing observed and computed concentrations, as shown in Figure 15 from Shieh et al. (1970). Observed and simulated values in the figure are both hourly averages, but those of Figure 16 (from Shir and Shieh, 1973) are a mixture of 2-hourly, 24-hourly, and 3-monthly averaged data. Scatter diagrams are good as they provide estimates of how well a model works at various observed concentration levels.

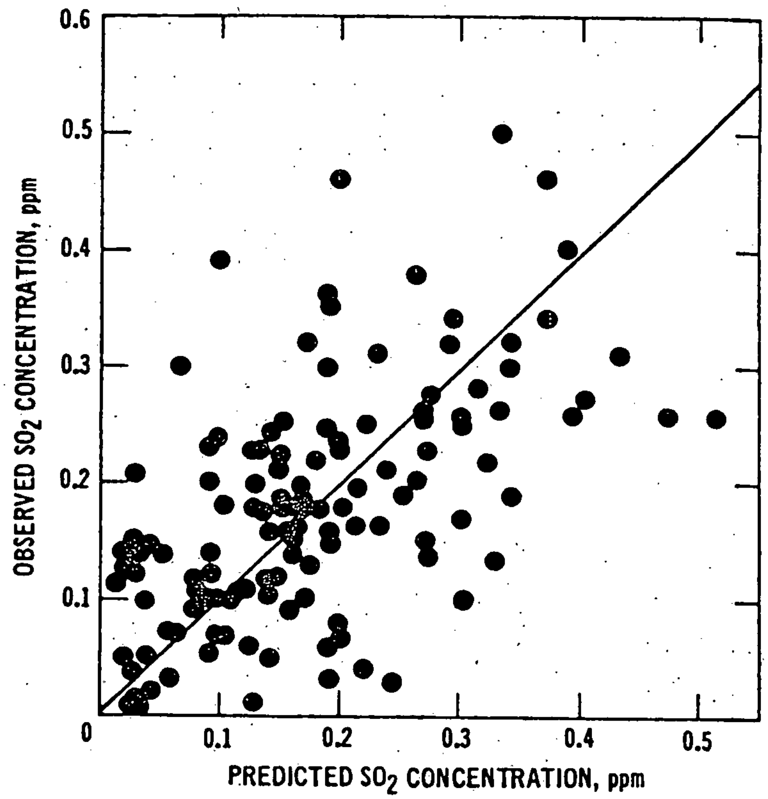


Figure 15. Plot of observed against predicted concentrations of SO_2 . (Taken from Shieh et al., 1970)

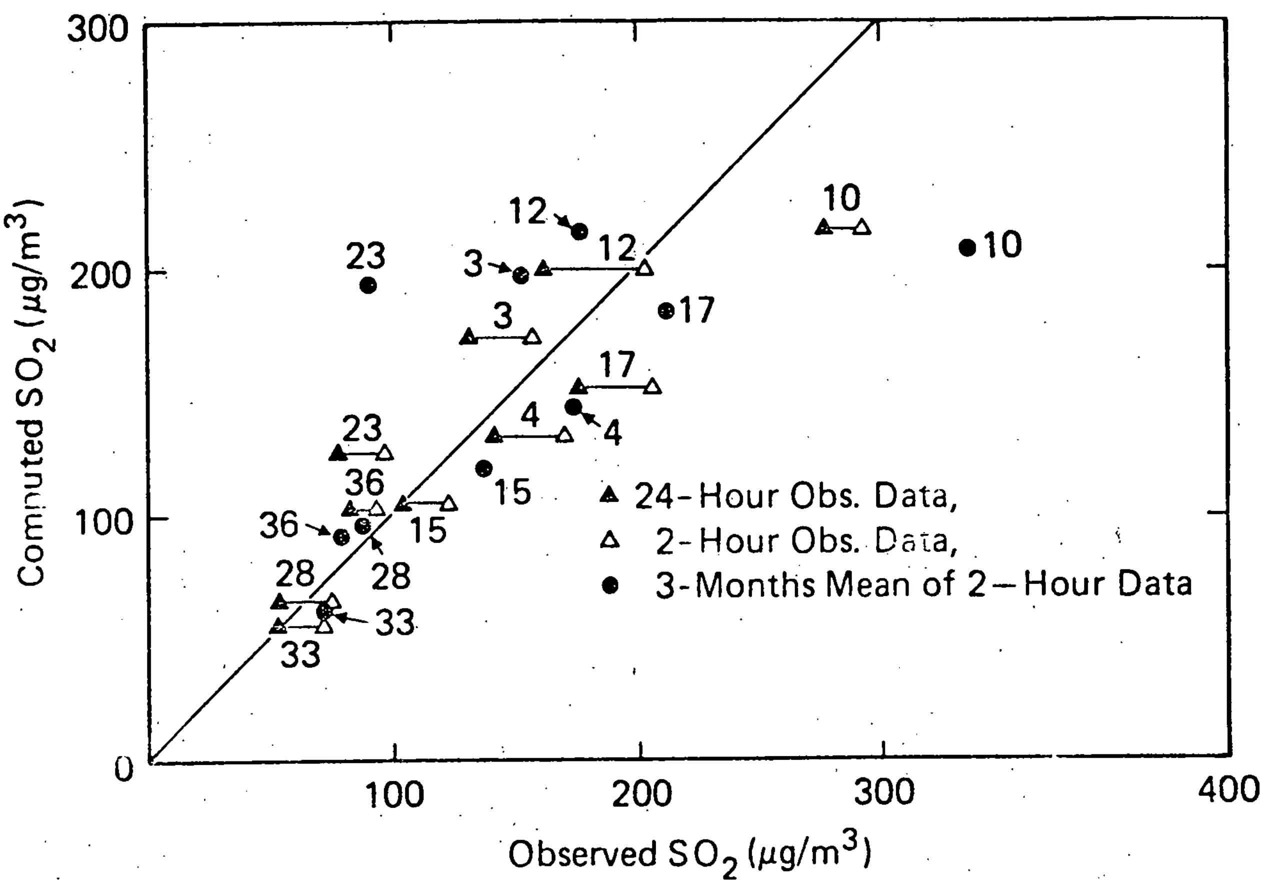


Figure 16. Comparison of observed vs. computed SO_2 concentration values (Feb. 1, 1965 to Feb. 26, 1965). Station numbers are plotted next to each data point. (Taken from Shir and Shieh, 1973)

To obtain insight on model accuracy as a function of meteorological conditions, it is possible to construct separate scatter diagrams for different meteorological regimes, e.g., for daytime unstable versus nighttime stable conditions or north versus south winds. In any case, scatter diagrams should be used in conjunction with the quantitative estimates of model accuracy discussed in Section C.

Concurrent cumulative frequency plots of observed and predicted concentrations have been used by Fortak (1970) to obtain qualitative estimates of model accuracy (Figure 17). When the "observed" curve lies above the "computed" curve, the model overpredicts concentration, and vice versa. Also shown at the bottom of the figure are frequency histograms of observed and predicted concentrations. This presentation is particularly good for quick identification of the frequency with which ambient air quality standards are exceeded.

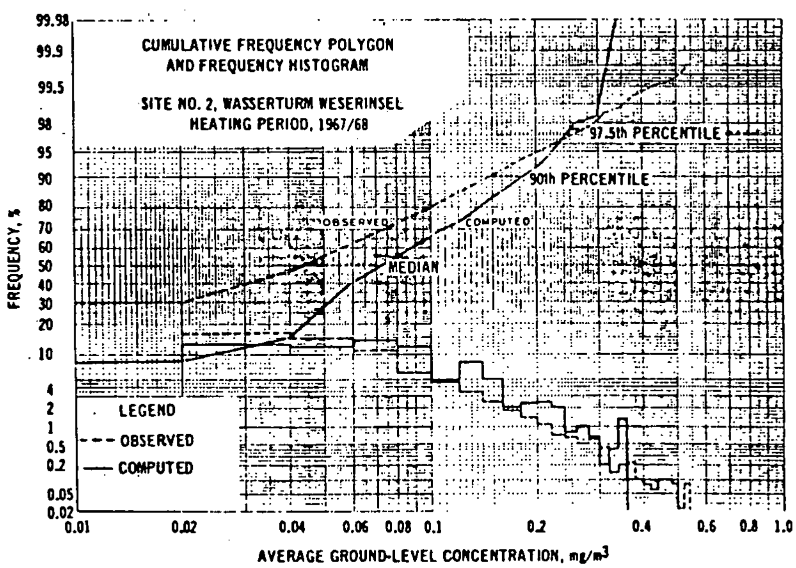


Figure 17. Comparison between observed and computed cumulative frequency plots of ground-level concentrations in downtown Bremen. (Taken from Fortak, 1970)

Cumulative frequency plots for concentrations valid over specified ranges of input meteorological parameters, as shown in Figure 18 from Shir and Shieh (1973), can reveal systematic differences between observed and computed values. Results show that the model overestimates (underestimates) concentrations when temperatures are low (high).

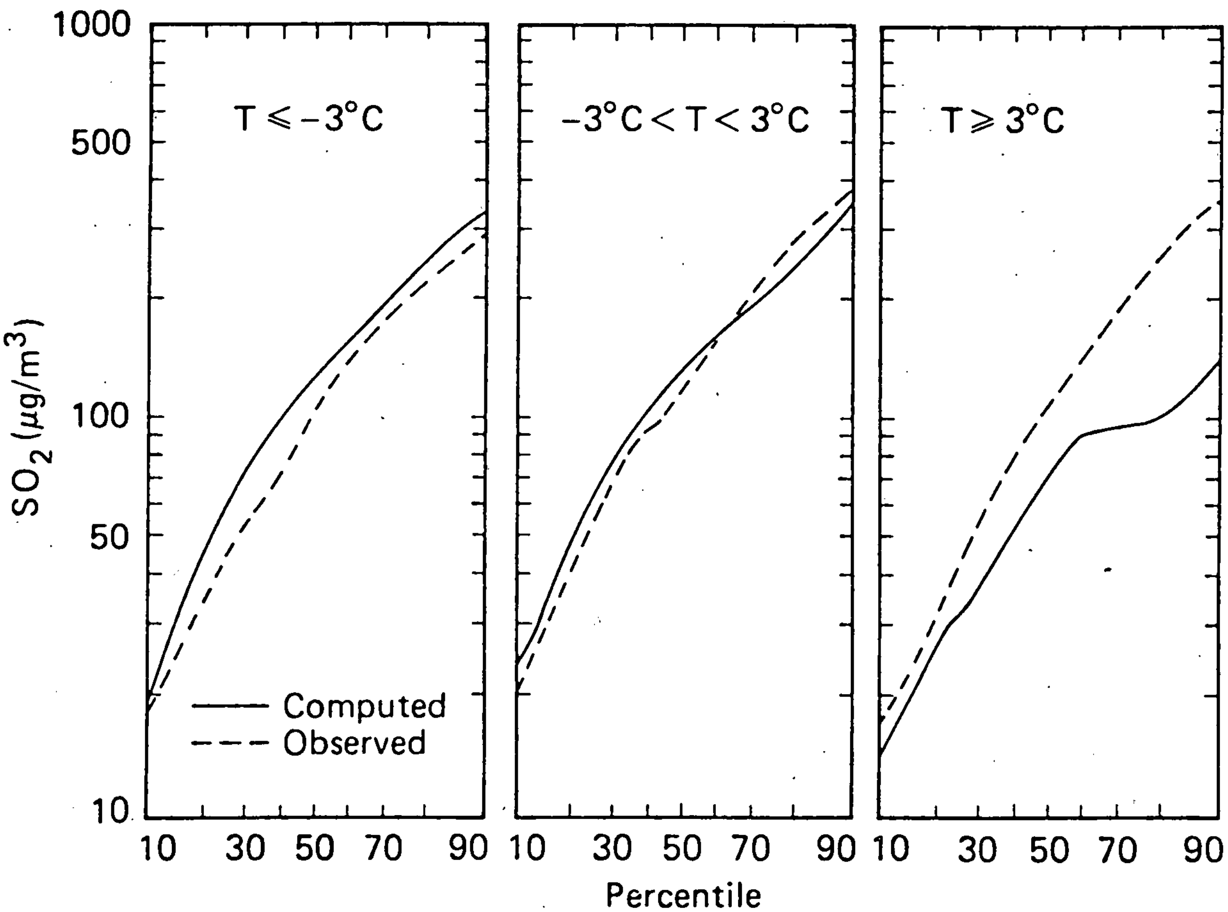


Figure 18. Comparison between observed and computed 2-hour averaged frequency distributions of SO_2 concentration according to temperature range. (Taken from Shir and Shieh, 1973)

The propensity of a model to under or overpredict can also be demonstrated by plotting the frequency distribution of the differences between observed and predicted concentrations. This was done by Tesche et al. (1979) for hourly averaged NO_2 and O_3 values (Figure 19) and results showed that most of the differences were within ± 8 pphm .

As the above technique does not provide information on model performance as a function of concentration, the data in the figure were used to construct Figure 20, in which the deviations are plotted against concentration. Results showed a bias towards underestimation of O_3 concentration at low concentration and a reverse bias for NO_2 .

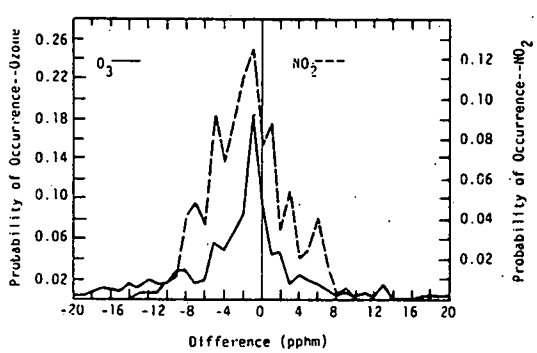


Figure 19. Deviation of calculated versus observed NO_2 and ozone concentrations from perfect correlation. (Taken from Tesche et al., 1979)

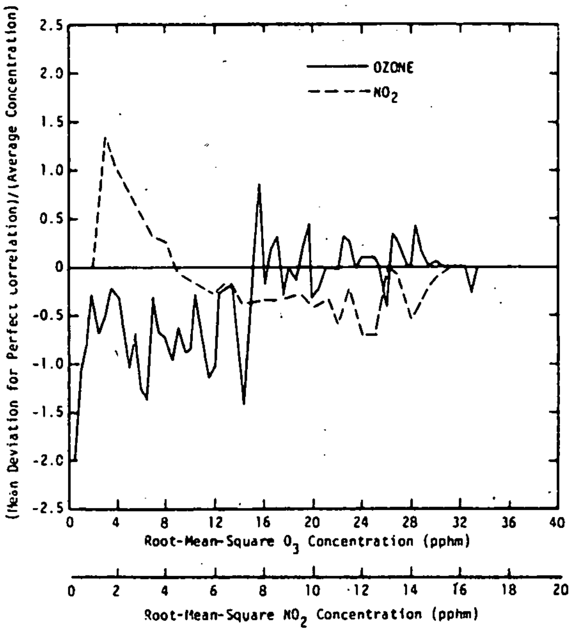


Figure 20. Normalized deviations about the perfect correlation line as a function of concentration level. (Taken from Tesche et al., 1979)

Systematic errors were also evident when Shieh and Shir (1976) plotted observed and predicted SO_2 values against air temperature (Figure 21). The concentrations due to area source emissions only (lower curve) parallels the upper (area plus point source emission) curve. Air temperature is not a direct parameter in the model formulation, but enters through parameterization of area source emission rates. Thus, it was concluded that the given functional relationship between area source emissions and temperature was not accurated for extreme temperatures.

To formulate a better area source emission algorithm, it was reasoned that area source emissions should depend on wind speed as well, as heat loss from a home increases as this parameter increases in value. When this

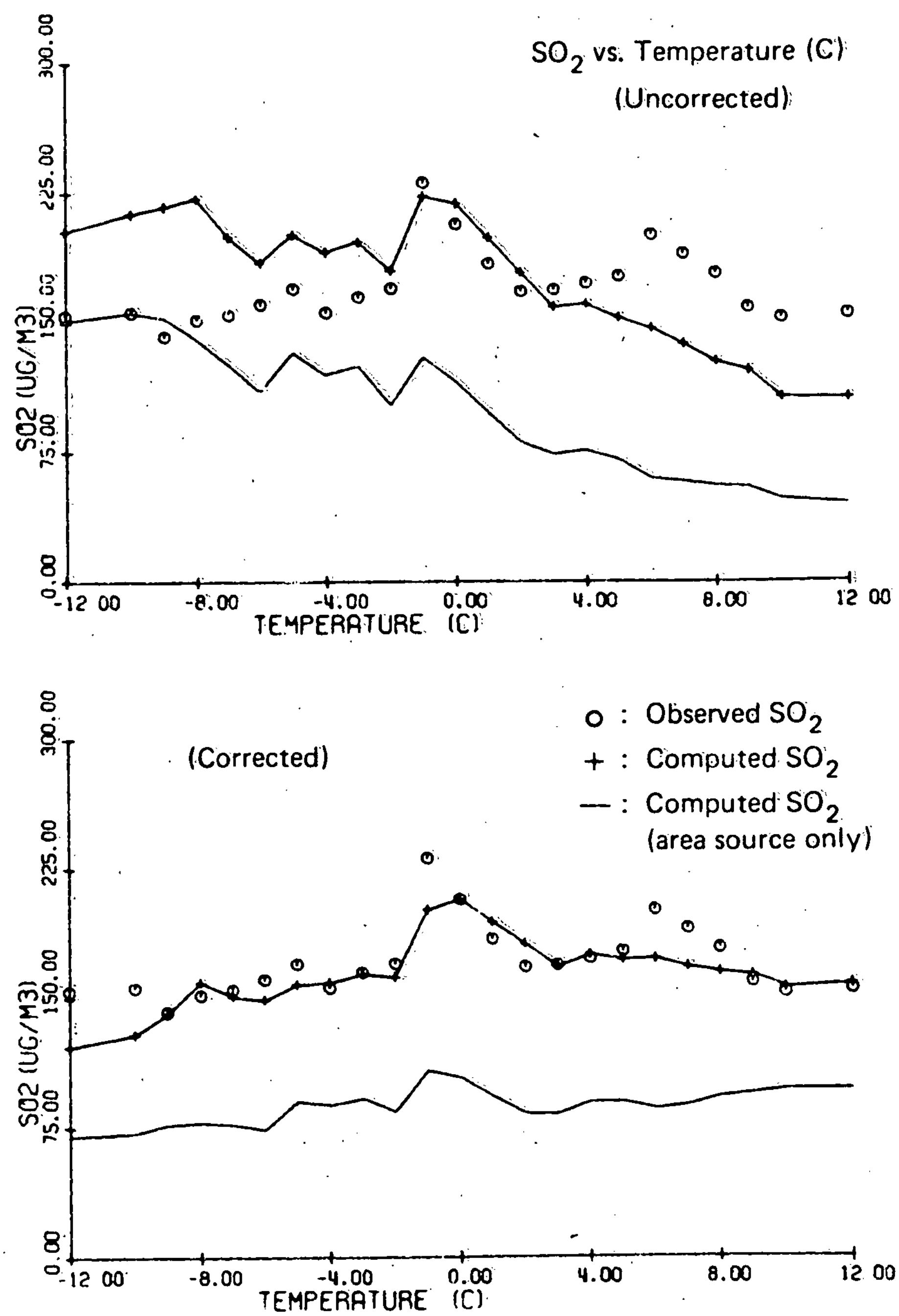


Figure 21. Computed and observed concentrations as a function of temperature before and after correction to remove systematic error. (From Shieh and Shir, 1976)

correction was applied to the area source emission formulation, the systematic errors disappeared as shown in the lower half of Figure 21. This approach appears to be useful in identifying and correcting causes of systematic errors in model formulation.

B. Tabular

In addition to the time series plots discussed in the previous section, Roberts et al. (1970) presented three types of tabular comparisons between observed and predicted concentrations. The summary statistics in Table 1 will be discussed in Section C, but of interest now is the tabular presentation of percentages of calculated values within various given "tolerances" (in ppm) of corresponding observed values at various sites for prediction periods ranging from 1 to 24 hours.

One strength of this type of presentation is that it demonstrates how well a model does over various forecast time intervals. A weakness of the approach is that it provides no insight as to how well the model does during particular situations, e.g., high versus low concentration periods.

STATISTICAL EVALUATION OF INTEGRATED PUFF MODEL DATA: PERCENTAGE OF CALCULATED POLLUTANT CONCENTRATIONS WITHIN DESIGNATED TOLERANCE LIMITS OF SAMPLE DATA[a]

Item	Hour avg.	TAM stations 1	2	3	4	5	1-5
Mean (μ_{obs})		0.06 ppm	0.12 ppm	0.33 ppm	0.15 ppm	0.06 ppm	0.14 ppm
$\frac{(\mu_{obs}-\mu_{calc})}{\mu_{obs}} \times 100\%$		- 10%	+ 16%	+ 14%	- 3%	- 9%	+ 7.4%
Std. dev. (obs-calc)	1	0.09 ppm	0.10 ppm	0.20 ppm	0.13 ppm	0.10 ppm	0.13 ppm
Std. dev/mean		1.7	0.88	0.62	0.84	1.6	0.93
% · 0.025 ppm	1	58%	37%	12%	26%	45%	35%
% · 0.05 ppm	1	69%	63%	23%	53%	71%	57%
% · 0.1 ppm	1	84%	81%	47%	79%	89%	77%
No. data points	1	288	576	432	408	624	2328
Std. dev. (obs-calc)	2	0.08	0.09	0.18	0.11	0.09	0.12
Std. dev/mean	2	1.3	0.78	0.56	0.77	1.5	0.83
% · 0.025 ppm	2	58%	40%	15%	29%	45%	37%
% · 0.05	2	70%	66%	27%	52%	70%	58%
% · 0.1 ppm	2	83%	83%	49%	81%	88%	78%
No. data points	2	144	288	216	204	312	1164
Std. dev (obs-calc)	6	0.07	0.08	0.14	0.08	0.07	0.09
Std. dev/mean	6	1.1	.65	0.42	0.58	1.2	0.64
% · 0.025 ppm	6	56%	36%	10%	32%	46%	36%
% · 0.05 ppm	6	67%	66%	29%	54%	72%	59%
% · 0.1 ppm	6	90%	88%	56%	85%	91%	83%
No. data points	6	48	96	72	68	104	388
Std. dev (obs-calc)	24	0.04 ppm	0.05 ppm	0.10 ppm	0.07 ppm	0.04 ppm	0.06 ppm
Std. dev/mean	24	0.71	0.44	0.30	0.45	0.70	0.43
% · 0.025 ppm	24	25%	46%	17%	59%	.58%	43%
% · 0.05 ppm	24	75%	62%	39%	77%	77%	66%
% · 0.1 ppm	24	100%	96%	67%	88%	96%	90%
No. data points	24	12	24	28	17	26	97

[a] Chicago, January, 1967

Table 1. Statistical summary of Roberts et al. (1970), where μ represents a mean value.

The second tabular presentation given by Roberts et al. was a bivariate frequency distribution of observed versus predicted concentrations in various concentration intervals (Table 2). Values in the squares along the positive diagonal (outlined in heavy lines) represent "successful predictions" and the technique overcomes the deficiency mentioned in conjunction with the previous tabular presentations, as it shows the success of the model at different observed concentration levels.

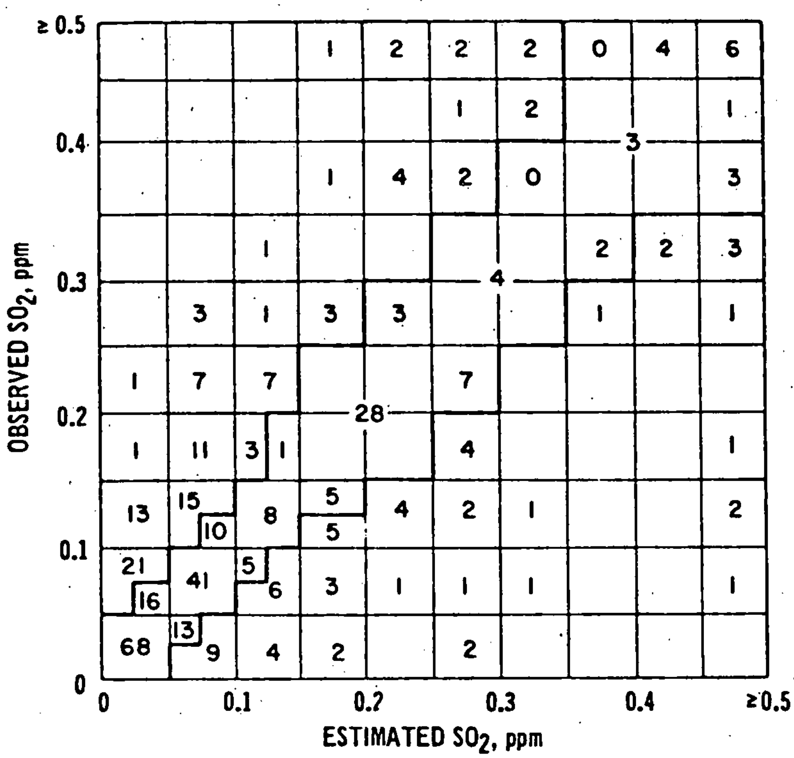


Table 2. Six-hour pollutant averages, Chicago, January 1967. (From Roberts et al., 1970)

The third tabular verification technique presented was a contingency table (Table 3) for estimating the accuracy in forecasting air pollution "incidents" for 6-hour and 24-hour forecast periods. Three levels of "incidents" are listed in the tables, i.e., at threshold concentrations of 0.1, 0.2, and 0.3 ppm, respectively.

Verification of model accuracy is based on the following criteria:

1) an "incident" is said to have occurred if the observed average concentration is greater than the threshold concentration less a tolerance (of 0.025 ppm), and

2) an "incident" is said not to have occurred if the observed average concentration is less than the threshold concentration plus a tolerance (of 0.025 ppm).

As shown in the tables, the model did quite well, as most incidents predicted to occur did, in fact, verify. The same is true for most of the times when incidents were forecast not to occur. Forecasted "incidents" are useful to public officials, who can evoke emergency actions to prevent their actual occurrence, e.g., they can force temporary change-overs to expensive low-sulfur fuels for periods when meteorological conditions are forecasted to be conducive to "incident" level concentrations.

FREQUENCY OF 24-HR INCIDENTS FOR STATIONS 1–5
JANUARY 1967

Symbol	Threshold	Tolerance	Skill score (Based on chance)
x	0.1 ppm	0.025	0.76
y	0.2 ppm	0.025	0.73
z	0.3 ppm	0.025	0.75

CONTINGENCY

		Predict	
		Yes	No
Occur	Yes	x 44 y 17 z 9	x 9 y 6 z 1
	No	x 3 y 3 z 3	x 41 y 71 z 84

FREQUENCY OF 6-HR INCIDENTS FOR STATIONS 1–5
JANUARY 1967

Symbol	Threshold	Tolerance	Skill score (Based on chance)
x	0.1 ppm	0.025	0.76
y	0.2 ppm	0.025	0.64
z	0.3 ppm	0.025	0.80

CONTINGENCY

		Predict	
		Yes	No
Occur	Yes	x 145 y 62 z 32	x 32 y 28 z 10
	No	x 14 y 20 z 6	x 197 y 278 z 340

Table 3: Contingency tables from Roberts et. al. (1970)

As with graphical summary techniques, if not even more so, tabular summary techniques should be presented in conjunction with the quantitative summary statistics described in the following section.

C. Summary Statistics

The most elementary quantitative estimate of how well a model predicts concentration is a comparison between average predicted concentration $\overline{C}_p$ and average observed concentration $\overline{C}_0$ as is done in Table 4 from Fortak (1970). Presentation of such values on a station-by-station basis, as is done in the table, gives a superior estimate of model accuracy to that obtained from presentations of similar values averaged over many sites in a given air basin, as average under and overpredictions at particular sites will cancel out in the latter procedure.

OBSERVED AND CALCULATED MEAN SO_2 CONCENTRATIONS IN BREMEN; HEATING PERIOD, 1967-1968
($mg\ m^{-3}$)

Site	1		2		3		4	
	Calc.	Obs.	Calc.	Obs.	Calc.	Obs.	Calc.	Obs.
November	0.14	0.11	0.15	0.10	0.14	0.08	0.05	0.10
December	0.10	0.08	0.12	0.08	0.10	0.06	0.03	0.07
January	0.10	0.10	0.12	0.13	0.09	0.08	0.04	0.08
February	0.09	0.08	0.12	0.08	0.07	0.06	0.04	0.08
March	0.08	0.07	0.10	0.04	0.07	0.05	0.04	0.07
April	0.09	0.07	0.11	0.05	0.06	0.05	0.05	0.07
May	0.05	0.06	0.07	0.03	0.04	0.03	0.03	0.05
Total	0.09	0.08	0.12	0.08	0.08	0.06	0.04	0.08

Table 4: Tabular results from Fortak (1970)

Results from the above study are for a fixed forecast period and give no information on model accuracy as a function of this parameter. However, Zannetti and Switzer (1979) analyzed the forecast accuracy of six statistical models for periods ranging from 1 to 8 hours (Figure 22). A "correct" forecast was determined from a three-by-three contingency table for high, average, and low concentrations.

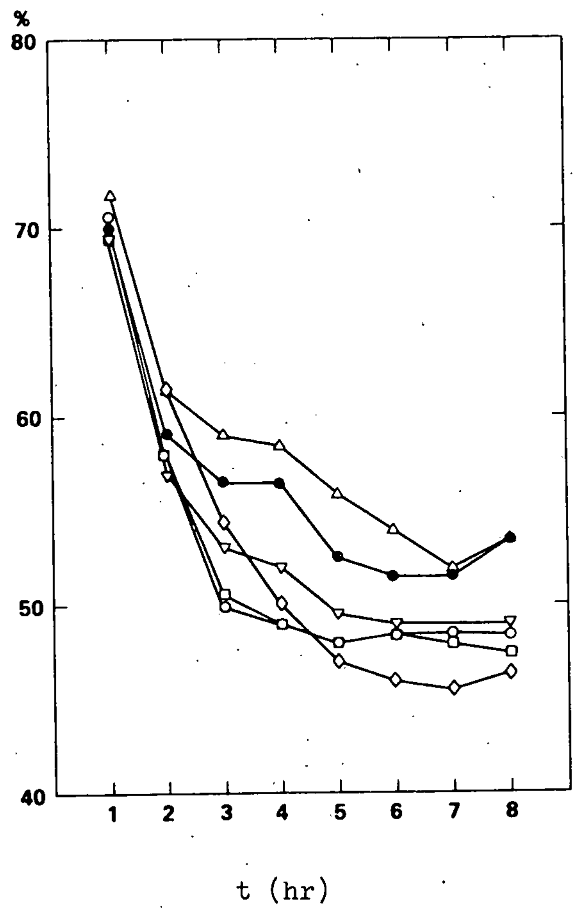


Figure 22. Percentage of "correct" predictions versus forecast time for six models. (Taken from Zannetti and Switzer, 1979)

However, even at a particular site, over and underpredicted values cancel out in the computation of $\overline{C}_p$, and thus Pasquill (1970) evaluated the root mean square predicted error σ_Δ from

(2) $$\sigma_\Delta = \sqrt{\frac{1}{N} \Sigma \, (C_0 - C_p)^2} \; .$$

Also evaluated was the coefficient of the variation of the predicted errors (or the relative root mean square error)

(3) $$R = \frac{\sigma_\Delta}{\overline{C}_0} \; ,$$

which gives a quantitative estimate of the relative error in σ_Δ. These values for the entire air basin, but not for each of the six sites in the network, are presented in Table 5 for three different models.

	Method of calculation	Mean of 6-hour concentration, $\mu g\ m^{-3}$ Obs. $\overline{C}$	Calc. $\overline{C}$	r.m.s.difference between calculated and observed C	r.m.s. ÷ obs. $\overline{C}$
All periods[a]	a. Wind fluctuation	68	38	84	1.2
	b. Broad estimates of spread (stability categories)	68	88	132	1.94
	c. Regression on temperature and wind	68	–	68	1.0
119 selected periods[b]	a. Wind fluctuation	66	53	63	0.95
	b. Regression on temperature and wind	66	–	57	0.86

[a] Methods a. and b. follow Pasquill[2]

[b] Steady wind direction and speed (6 a.m. to 12 noon and 12 noon to 6 p.m. only) for which most confident estimates of emission were made.

Table 5: Statistical summary from Pasquill (1970)

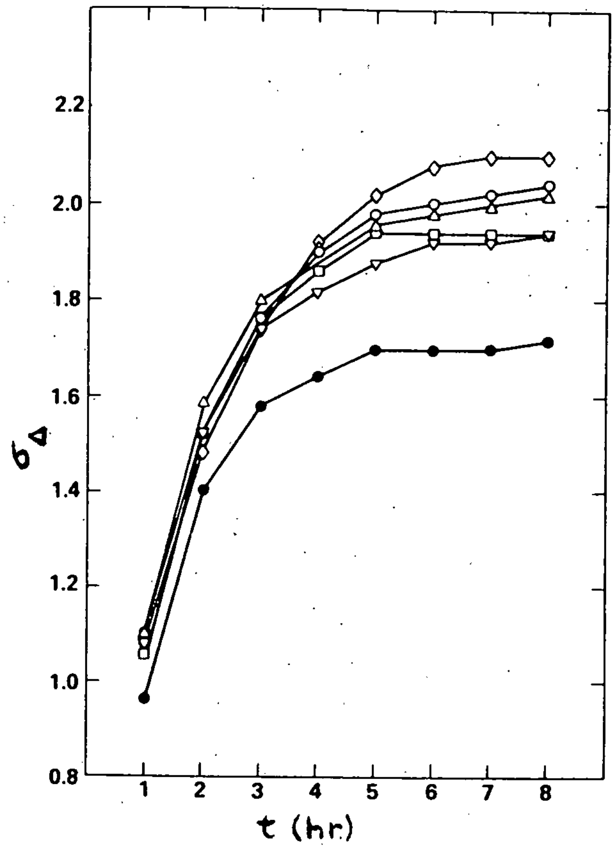


Figure 24. Root mean square of the prediction error plotted against the forecast time for six models for high concentration values. (Taken from Zannetti and Switzer, 1979)

Computed values of $\overline{C}_0$ and $\overline{C}_p$ were used by Roberts et al. (1970) to compute the average fractional percent error $\overline{E}_p$ from

(4)
$$\overline{E}_p\ (\%) = \frac{\overline{C}_0 - \overline{C}_p}{\overline{C}_0} \times 100\%$$

on a station-by-station basis (Table 1). This parameter estimates the relative difference between $\overline{C}_0$ and $\overline{C}_p$.

The root mean square error of six statistical forecast models studied by Zannetti and Switzer (1979) as a function of forecast time is shown in Figure 23. To demonstrate model accuracy as a function of concentration level, similar results are shown in Figure 24 for high concentration "episode" periods.

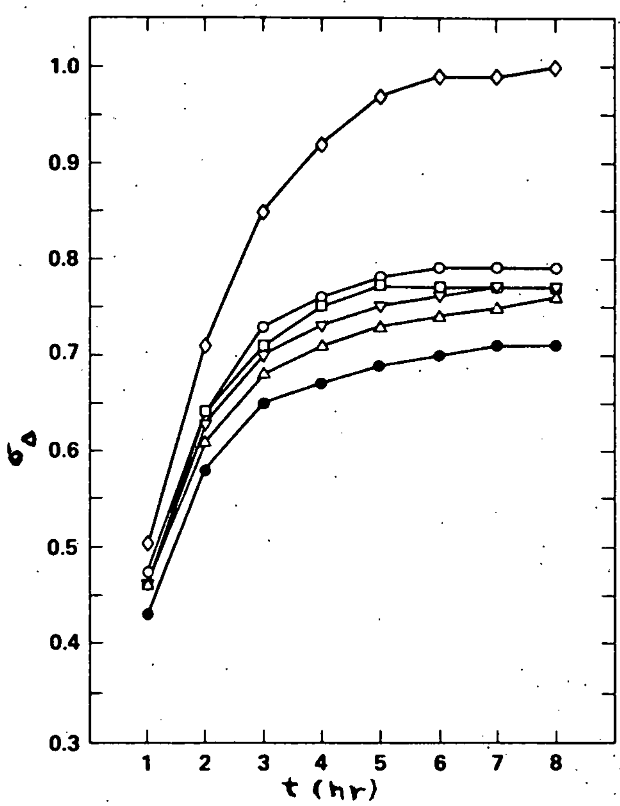


Figure 23. Root mean square of the prediction errors plotted against forecast time for six forecasting models. (Taken from Zannetti and Switzer, 1979)

Fractional errors E given by

(5) $$E = \frac{C_0 - C_p}{C_0}$$

were used by Hilst (1970) to evaluate the average and root mean square values of this parameter from

(6) $$\overline{E} = \left(\overline{\frac{C_0 - C_p}{C_0}}\right)$$

and

(7) $$\sigma_E = \sqrt{\frac{1}{N}\sum\left(\frac{C_0 - C_p}{C_0}\right)^2 - \left(\overline{\frac{C_0 - C_p}{C_0}}\right)^2} .$$

This last equation was incorrectly given (due to a typographical error) as

(8) $$\sigma_E = \sqrt{\frac{1}{N}\sum\left(\frac{\overline{C_0 - C_p}}{C_0}\right)^2 - \left(\frac{C_0 - C_p}{C_0}\right)^2} .$$

Results were presented as C_0 versus E (Figure 25), with $\overline{E}$ and σ_E having values of -0.642 and 1.225 , respectively.

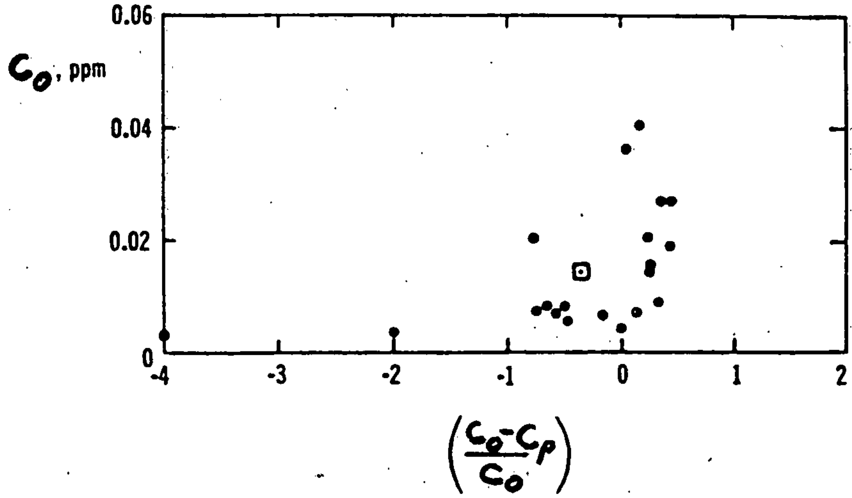


Figure 25. Joint values of observed values of SO_2 concentrations and fractional error in values predicted by TRC model for period 0600-0800, October 30, 1968. (Taken from Hilst, 1970)

A better estimate of the average error than that given in Equation (6) for the average fractional error $\overline{E}$ is the average absolute fractional error $\overline{|E|}$ given by

(9)
$$\overline{|E|} = \overline{\left|\frac{C_0 - C_p}{C_0}\right|} .$$

This parameter is superior to $\overline{E}$ as an estimate of model accuracy, as overpredictions and underpredictions do not cancel each other out in its computation. A similar argument holds for the superiority of the absolute average error $|\Delta C|$ given by

(10)
$$|\Delta C| = \overline{|C_0 - C_p|}$$

as compared to the average error ΔC given by

$$\Delta C = \overline{(C_0 - C_p)} \quad . \tag{11}$$

Scatter diagrams, as described above, are frequently presented with quantitative estimates of model accuracy, e.g., Shieh et al. (1970) included values of $\bar{C}_0$, $\bar{C}_p$, and the overall model standard error of the estimate S , using a least-square regression of C_p on C_0 , where

$$S = \sqrt{\left(\frac{N-1}{N-2}\right)(\sigma_p^2 - b^2\ \sigma_0^2)} \tag{12}$$

where

$$b = \frac{\Sigma C_0\ C_p - N^{-1}\ (\Sigma C_0)(\Sigma C_p)}{(\Sigma C_0^2) - N^{-1}\ (\Sigma C_0)^2} \quad . \tag{13}$$

For the data given in Figure 15, $\bar{C}_0$, $\bar{C}_p$, and S had values of 0.19, 0.18, and 0.09 ppm, respectively.

Estimates of the overall linear correlation coefficient r between C_0 and C_p was computed by Shir and Shieh (1973) from

$$r = \frac{\Sigma[(C_0 - \bar{C}_0)(C_p - \bar{C}_p)]}{\sqrt{\Sigma[(C_0 - \bar{C}_0)^2]\ \Sigma[(C_p - \bar{C}_p)^2]}} \quad . \tag{14}$$

Results are shown in Table 6 on a station-by-station basis for input wind fields analyzed by objective and subjective methods.

TABLE 2. Correlation Coefficients of Computed vs. Observed 24-Hour Averaged SO_2 Concentrations

Station No.	Objective analyzed windfield		Turner's analyzed wind field	
	Linear Scale	Log Scale	Linear Scale	Log Scale
3	0.380	0.334	0.398	0.325
15	0.617	0.598	0.635	0.620
17	0.363	0.466	0.408	0.502
23	0.471	0.647	0.444	0.642
33	0.832	0.725	0.830	0.718
4	0.177	0.284	0.220	0.313
10	0.139	0.145	0.105	0.125
12	0.830	0.821	0.812	0.804
28	0.828	0.886	0.823	0.878
36	0.914	0.910	0.937	0.940
TOTAL	0.654	0.806	0.659	0.809

Table 6. Correlation coefficients of computed vs. observed 24-hour averaged SO_2 concentrations from Shir and Shieh (1973)

Since observed concentration data frequently possess a log normal (rather than a normal) distribution, logarithmic correlation coefficients r' were also computed from

$$r' = \frac{\Sigma[(\ln C_0 - \overline{\ln C_0})(\ln C_p - \overline{\ln C_p})]}{\overline{\Sigma[(\ln C_0 - \overline{\ln C_0})^2]\ \Sigma[(\ln C_p - \overline{\ln C_p})^2]}} \quad (15)$$

Correlating the logarithm of observed data having a log normal distribution reduces the range of the values and thus leads to better correlations as shown in the table.

To demonstrate model validity as a function of forecast time, Zannetti and Switzer (1979) computed correlation coefficients for six models as a function of forecast period (Figure 26).

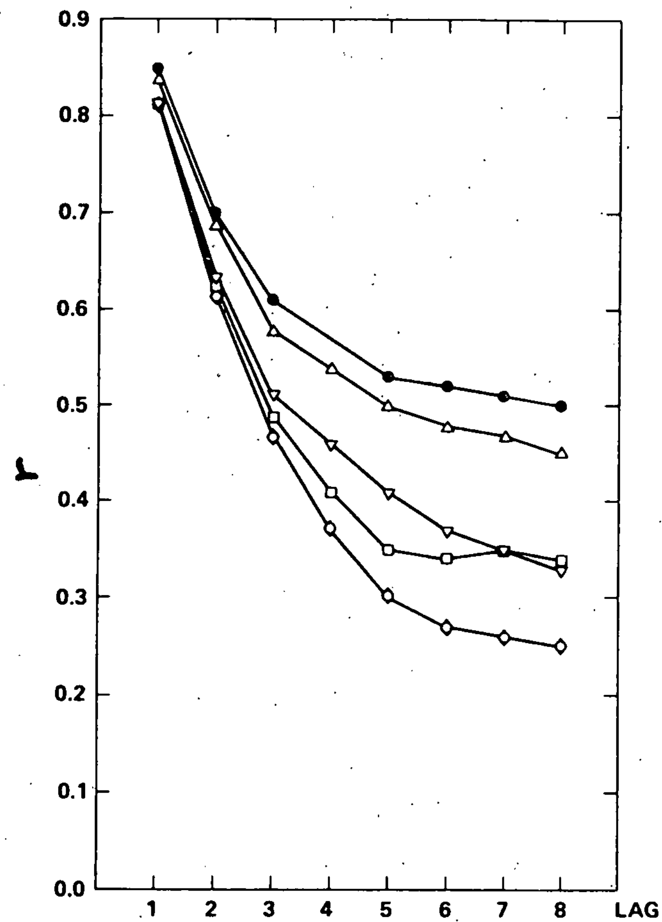


Figure 26. Correlation coefficient between observed and predicted values plotted against the forecast time for six models. (Taken from Zannetti and Switzer, 1979)

One problem with correlation coefficients is that they only indicate whether or not the "trend" is properly simulated. They do not provide a quantitative estimate of the magnitude of the error, e.g., if the model always overpredicts concentrations by a constant factor, the correlation will be perfect. Thus correlation coefficients must be used in conjunction with quantitative estimates of the magnitude of differences between C_0 and C_p. This was done by Shieh and Shir (1976), who presented (Table 7) concurrent values of the following at various sites: $\overline{C}_0$, $\overline{C}_p$, σ_0, σ_p, and r.

Station Number	No. of Samples	Average		Standard Deviation		Correlation Coefficient
		Observed	Computed	Observed	Computed	
3	295	155.8	192.3	126.9	124.2	0.478
15	232	121.5	114.2	115.1	104.2	0.416
17	290	204.4	157.5	110.9	119.7	0.400
23	267	96.2	159.4	110.9	153.8	0.479
33	235	71.7	66.1	81.2	83.3	0.687
4	268	169.7	160.4	158.4	97.9	0.388
10	288	315.7	215.9	225.4	137.5	0.099
12	275	201.4	215.8	158.3	164.4	0.559
28	282	76.1	94.4	107.2	137.1	0.713
36	286	91.5	126.7	89.6	140.2	0.579
Station Average	300	153.8	153.0	70.8	70.1	0.438
Total Sample	2,718	152.8	152.4	153.5	137.5	0.472

Table 7. General result of model computations of 2-hr. average SO_2 concentration from Shieh and Shir (1976)

They also evaluated the normalized root mean square error ϵ from

$$\epsilon = \frac{\sigma_\Delta}{\sigma_0} , \tag{16}$$

where σ_0 is the standard deviation of the observed concentrations and σ_Δ is given by Equation (2) . This parameter relates the variance of the prediction error to the variance in the observed concentrations.

III. Conclusion

This paper has surveyed some of the techniques that have been used by meteorologists to validate air pollution prediction models. Characteristics desired in any validation program include a demonstration of how well the model validates: 1) over the entire range of observed concentrations; 2) under various meteorological and source strength conditions; 3) for both area and point sources; and 4) for different forecast periods.

Graphical summaries should demonstrate qualitatively how well the model validates: 1) over time at a single station; and 2) in space over an entire air basin at horizontal levels at the surface and aloft, as well as in vertical cross sections.

Quantitative estimates of the magnitude of the average absolute error should be given, as opposed to that of the average error, as overestimates and underestimates can cancel out. These estimates should be given for each site, as opposed to one value for the entire air basin, for the same reason.

Data could be smoothed, but logarithmic plots should be avoided as they visually distort the magnitude of the errors. Attempts should be made to identify and correct for systematic errors in input meteorological and source strength data.

While no one technique possesses all of the above characteristics, a combination of qualitative and quantitative techniques should be used. Qualitative presentations possessing some of the above characteristics include time series plots, transections, isopleth analysis (including those for ΔC), scatter diagrams, cumulative frequency plots, and bivarial frequency distributions. Quantitative presentations that could be used with the above qualitative techniques include mean values, standard deviations, standard error of the estimates, and correlation coefficients.

Undoubtedly there is much that air pollution meteorologists can learn from statisticians about existing verification techniques. In addition, the more statisticians become aware of the practical problems associated with air pollution modeling, the more they will be able to develop new techniques to overcome these problems.

One of the most important practical problems facing air pollution meteorologists is in the area of network design. Important questions to be answered include how many observation sites are required, where should they be placed, and how frequently should they be sampled? Another important practical problem arises because observations are point values and predicted values are volume-averaged. Two aspects of this problem involve the effect of emission grid size on the magnitude of predicted values and comparison of predicted roof top values with observed urban canyon values.

Given these problems, and others associated with model formulation, such as numerical diffusion and approximations in the basic meteorological and photochemical equations, it is probably unrealistic to expect perfect agreement between observed and predicted values. Statisticians may thus be able to help estimate the limits to the accuracy of air pollution prediction models.

This paper has outlined one area in which fruitful collaboration between air pollution meteorologists and statisticians can occur. Others exist in air pollution meteorology, such as statistical forecasting methods, and there are many additional ones in other branches of meteorology, such as the verification of weather modification projects. Hopefully, this paper is a step to a fruitful collaboration between statisticians and meteorologists.

REFERENCES

Dietzer, B. (1976), Preliminary Results of the Application of a Statistical Model on Interregional Transport of SO_2 , Reprint from 7th International Technical Meeting on Air Pollution Modeling and its Application, September 7-10, 1976, Airlie, Virginia, 14 pp.

Fortak, H. G. (1970), Numerical Simulation of the Temporal and Spatial Distributions of Urban Air Pollution Concentration, Proceedings of Symposium on Multiple-Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 9.1-9.34.

Hilst, G. R. (1970), Sensitivities of Air Quality Prediction to Input Errors and Uncertainties, Proceedings of Symposium on Multiple-Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 7.1-7.24.

Johnson, W. B., et al. (1970) Development of a Practical, Multi-Purpose Urban Diffusion Model for Carbon Monoxide, Proceedings of Symposium on Multiple-Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 5.1-5.38.

Johnson, W. B., et al. (1971), Field Study for Initial Evaluation of an Urban Diffusion Model for Carbon Monoxide, Stanford Research Institute Report, Project #8563, 144 pp.

Lantz, R. B., et al. (1976), Diffusion Model Comparisons to Measured Data in Complex Terrain, Preprint Volume 3rd AMS Symposium on Atmospheric Turbulence, Diffusion, and Air Quality, October 19-22, 1976, Raleigh, NC, 476-483.

Pasquill, F. (1970) Prediction of Diffusion Over an Urban Area - Current Practice and Future Prospects, Proceedings of Sympsosium on Multiple-

Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 3.1-3.26.

Roberts, J. J., et al. (1970), An Urban Atmospheric Dispersion Model, Proceedings of Symposium on Multiple-Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 6.1-6.72.

Shieh, L. J., et al. (1970), A Model of Diffusion in Urban Atmospheres: SO_2 in Greater New York, Proceedings of Symposium on Multiple-Source Urban Diffusion Models, October 27-30, 1969, Research Triangle Park, NC, 10.1-10.39.

Shieh, L. J., and Shir, C. C. (1976), Evaluation of an Urban Air Quality Model with Respect to Input Parameters, IBM Technical Report #6320-3351, August 1976, 48 pp.

Shir, C. C., and Shieh, L. J. (1973), A Generalized Urban Air Pollution Model and its Application to the Study of SO_2 Distributions in the St. Louis Metropolitan Area, J. Appl. Meteor., 13 (2), 185-204.

Slade, D. H. (1968), Meteorology and Atomic Energy, USAEC Report, available from NTIS, 446 pp.

Stephens, T., and McCaldin, R. O. (1971), Attenuation of Power Station Plumes as Determined by Instrumented Aircraft, Environ. Sci. and Tech., 5 (7), 615-621.

Tesche, T. W., et al. (1979), Recent Verification Studies with the SAI Urban Airshed Model in the South Coast Air Basin, Preprint Volume Fourth Symposium on Turbulence, Diffusion and Air Pollution, January 15-18, 1979, Reno, Nevada, 307-312.

Zannetti, P., and Switzer, P. (1979), Some Problems of Validation and Testing of Numerical Air Pollution Models, Preprint Volume Fourth Symposium on Turbulence, Diffusion and Air Pollution, January 15-18, 1979, Reno, Nevada, 405-410.