STATISTICAL EVALUATION OF AIR QUALITY SIMULATION MODELS

by

K. P. MacKay
San Jose State University

and

R. D. Bornstein
San Jose State University
and
Stanford University

TECHNICAL REPORT NO. 56
APRIL 1982

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Statistical Evaluation
of Air Quality Simulation Models

K. P. MacKay* and R. D. Bornstein**

Abstract. Quantitative and qualitative statistical performance techniques used in the evaluation of air quality simulation models are reviewed. These techniques are used to evaluate the difference and correlation between observed and predicted air pollution concentrations.

Various sources of error, including model and instrument errors, are reviewed. In addition, possible improper uses of statistical techniques are reviewed, as these can lead to erroneous estimates of the actual error.

1. Introduction. Evaluation of air quality simulation models can consist of up to four phases: (1) a qualitative scientific evaluation as to whether model assumptions are reasonable and consistent with basic physical principles; (2) a one time tuning of model constants; (3) a sensitivity analysis of model dependence on input parameters; and (4) an operational evaluation for particular model applications.

Early evaluation efforts were reviewed by Hillyer et al. (1979), while previously utilized individual performance measures were summarized by Benca and Seinfeld (1979), who also described a new assessment package called AQMAAP. Attributes of good model performance were identified and related to particular air quality issues by Hayes (1979).

The evaluation protocol developed for the Electric Power Research Institute (Bowne, 1980) included both operational evaluation and a sequential sensitivity analysis of each model module (e.g., transport, diffusion, and transformation). The performance assessment protocol of Ruff et al. (1980) contained loss functions for various types of poor model performance, while the protocol of Martinez et

*Department of Meteorology, San Jose State University, San Jose, CA 95192
**Department of Statistics, Stanford University, Stanford, CA 94305
al. (1980) discussed combining various performance measures into one "figure of merit". Such a figure could be either the worst or average evaluation statistic.

The recent American Meteorological Society (AMS) Workshop on Dispersion Model Performance (Fox, 1980, 1981) brought together model developers and users, regulatory personnel, and statisticians to develop an evaluation protocol for EPA. The resulting protocol included quantitative and qualitative techniques to evaluate the magnitude of the differences and correlations of the following observed and predicted concentrations (arranged in decreasing order of stringency): (1) paired (in time and space) concentrations; (2) paired (in various degrees of time and/or space) maximum concentrations; and (3) unpaired (in time) concentration cumulative frequency distributions.

Maximum (highest or second highest) concentrations are important in many health and regulatory situations. When comparing maximum observed and predicted concentrations, the following aspects are important: magnitude, downwind location (distance and direction), and time of occurrence (MacKay, 1980).

The current paper reviews quantitative and qualitative performance techniques recommended by the AMS Workshop, and reviews various sources of error and problems associated with the evaluation of air quality simulation models.

2. Quantitative Performance Measures. The AMS Workshop suggested quantitative estimates be made of the magnitudes of the difference and correlation between observed concentrations \( C_o \) and predicted concentrations \( C_p \), as well as between maximum observed concentrations \( (C_m)_o \) and maximum predicted concentration \( (C_m)_p \).

Four possible difference functions can be defined from \( (C_m)_o \) and \( (C_m)_p \) as shown in Table 1 (from Fox, 1980). The most stringent of these, \( \Delta C_m(L,t) \), is computed from the difference between the observed maximum concentration at time \( t \) and the predicted concentration at that site \( L \) at that time. The next two are moderately stringent. The parameter \( \Delta C_m(\Delta L,t) \) is computed as the difference between the observed and predicted maximum concentrations at time \( t \), with these concentrations not necessarily at the same sites. The parameter \( \Delta C_m(L,\Delta t) \) is computed as the difference between the observed maximum concentration and the predicted maximum concentration at that site, with these values not necessarily occurring at the same time. The final parameter \( \Delta C_m(\Delta L,\Delta t) \) is the least stringent as it is computed from observed and predicted maximum concentrations occurring at any site and time.
TABLE 1

Possible difference functions defined using observed and predicted maximum concentrations paired in time t at location L. The quantities $\Delta t$ and $\Delta L$ represent unpaired arguments.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Paired in:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta C_m (L, t)$</td>
<td>time and location</td>
</tr>
<tr>
<td>$\Delta C_m (\Delta L, t)$</td>
<td>time only</td>
</tr>
<tr>
<td>$\Delta C_m (L, \Delta t)$</td>
<td>space only</td>
</tr>
<tr>
<td>$\Delta C_m (\Delta L, \Delta t)$</td>
<td>neither time or space</td>
</tr>
</tbody>
</table>

2.1. Difference. The concentration difference function $\Delta (\cdot)$ or residual is defined by

\[
\Delta (\cdot) = (\cdot)_o - (\cdot)_p .
\]

The residual is used in computation of the following quantitative estimates of difference: bias, noise, and gross error.

A model is biased if it consistently over or under predicts in some or all situations. The bias of the paired concentrations or the selectively paired maximum concentrations can be estimated by the average residual $\bar{\Delta (\cdot)}$ given by

\[
\bar{\Delta (\cdot)} = \frac{1}{N} \sum\Delta (\cdot) ,
\]

where $N$ is the number of observations.

The bias of the unpaired frequency distributions can be estimated by use of $t$, z, Wilcoxon, or Mann-Whitney tests (Fox, 1980). The first two of these tests are parametric, while the latter two are nonparametric.

Noise is the root mean square error (RMSE) evaluated around the bias. For the paired concentrations and the selectively paired concentrations, it is estimated by the standard deviation of the residuals $\sigma_{\Delta}$ using

\[
\sigma_{\Delta} = \left( \frac{1}{N-I} \left[ \Delta (\cdot) - \bar{\Delta (\cdot)} \right]^2 \right)^{1/2}
\]

while the noise of the unpaired frequency distributions can be evaluated by use of the parametric $F$ or $\chi^2$ tests.
Gross error is estimated by measures in which over and under predictions do not cancel. For the paired concentrations and the selectively paired maximum concentrations, it can be estimated by either the RMSE error evaluated around zero $\sigma$ or the average absolute error $|\overline{\Delta(c)}|$, given respectively by

\begin{equation}
\sigma = \left( \frac{1}{N} \sum \Delta(c_j)^2 \right)^{1/2}
\end{equation}

and

\begin{equation}
|\overline{\Delta(c)}| = \frac{1}{N} \sum \left| c_j - \bar{c}_j \right|
\end{equation}

while the gross error of the unpaired frequency distributions can be evaluated by use of the nonparametric $\chi^2$ or Kolmogorov-Smirnov tests (Fox, 1980).

Bias (first moment of a distribution) is related to noise (second moment) and gross error by

\begin{equation}
\sigma \approx \left\{ \left[ \overline{\Delta(c)} \right]^2 + \sigma^2 \right\}^{1/2},
\end{equation}

where the approximation arises from the $N-1$ in (3). All information concerning normal distributions is in the bias and noise. In addition, for normally distributed variables, the bias has a normal distribution, the noise a chi-squared distribution, and the gross error a bivariate or compound distribution.

Gross error can be given by

\begin{equation}
\sigma = (\overline{\Delta(c)})^2 + \sigma_o^2 + \sigma_p^2 - 2r \sigma_o \sigma_p
\end{equation}

where $\sigma_o$ and $\sigma_p$ are the standard deviations of the observed and predicted concentrations, respectively, and $r$ is the correlation coefficient between the observed and predicted concentrations. Brier (1975) pointed out that (7) implies that the following must be satisfied to have perfect magnitude prediction

\begin{align}
\overline{\Delta(c)} &= 0 \\
\sigma_o &= \sigma_p \\
r &= 1
\end{align}

2.2. Correlation. Evaluation of the correlation between observed and predicted concentration values involves evaluation of temporal
correlation and/or spatial alignment determined using temporal or spatial correlation coefficients (Table 2).

**TABLE 2**

Correlation coefficients for paired concentration data and selectively paired maximum concentrations (from Fox, 1980). See text for explanation of entries.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Temporal</th>
<th>Spatial</th>
<th>Space-Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta C$</td>
<td>$r(\tau)$</td>
<td>$r_s$</td>
<td>$r$</td>
</tr>
<tr>
<td>$\Delta C_m(L,t)$</td>
<td>$r_m(\tau)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta C_m(\Delta L,t)$</td>
<td>$r_m(\tau,\Delta L)$</td>
<td>$\Delta D(\Delta L)$</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta C_m(L,\Delta t)$</td>
<td>$\Delta t_m$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta C_m(\Delta L,\Delta t)$</td>
<td>$\Delta t_m$</td>
<td>$\Delta D(\Delta L,\Delta t)$</td>
<td>-</td>
</tr>
</tbody>
</table>

Three types of correlation coefficients can be calculated using observed and predicted paired concentration values. The cross-correlation (temporal) coefficient $r(\tau)$ is calculated using values at one site at all positive and negative time lags $\tau$ from the general definition of the correlation coefficient. The spatial correlation coefficient $r_s$ is calculated using values at a zero time lag at a given time using data from all sites together. The Pearson correlation coefficient $r$ is calculated using all values at a zero time lag for all times and for all sites together.

Given the definitions of the maximum concentration difference functions in Table 1, it is only appropriate to calculate correlation coefficients for the temporal case for $\Delta C_m(L,t)$ and $\Delta C_m(\Delta L,t)$ according to Fox (1980). As shown in Table 2, $r_m(\tau)$ and $r_m(\tau,\Delta L)$ use the same maximum concentration data used to evaluate $\Delta C_m(L,t)$ and $\Delta C_m(\Delta L,t)$, respectively.

The above statistics can be used to define the time lag $\tau$ that yields the highest correlations. However, zero time lags are generally used in evaluation studies.

Time and space differences can be used in some of the cases where correlation coefficients are inappropriate. The time difference between observed and computed maximum concentrations $\Delta t_m$ (Table 2)
can be used with the data utilized in evaluating $\Delta C_m(L, \Delta t)$, i.e., data from a single site, but at any time. When $\Delta t_m$ is computed using data associated with $\Delta C_m(\Delta L, \Delta t)$, it is calculated with data from anywhere in the region and at any time.

Instead of spatial correlation coefficients, vector differences (of distance and direction) between locations of observed and predicted maximum concentrations are used. When evaluated in conjunction with data associated with $\Delta C_m(\Delta L, t)$, it is determined using maximum predicted and observed concentrations at a given time anywhere in the region, but with data associated with $\Delta C_m(\Delta L, \Delta t)$, it is evaluated with data from anywhere in the temporal and spatial fields.

Correlation coefficients are generally evaluated in conjunction with a linear regression analysis in which

\begin{equation}
C_o = b_0 + b_1 C_p,
\end{equation}

where $b_0$ is the intercept and $b_1$ is the slope, given by

\begin{equation}
b_1 = r \frac{\sigma_o}{\sigma_p}.
\end{equation}

Thus, for perfect correlation it is necessary that (8b) and (8c) be satisfied, as well as

\begin{equation}
b_0 = 0
\end{equation}

and

\begin{equation}
b_1 = 1
\end{equation}

3. Qualitative Analysis Techniques. Qualitative analysis techniques such as histograms, cumulative frequency plots, isopleth analysis, transections, and scatter diagrams can provide information on differences and/or correlation.

3.1. Difference. Histograms and cumulative frequency plots give information on differences between observed and predicted concentrations. Specific data sets to be used in these analyses are shown in Table 3.

A frequency distribution plot of the distance (in number of grid cells) from each observation site at which predicted concentrations first equaled observed values was used by Tesche et al. (1979).
TABLE 3

Quantities to be summarized in histograms and cumulative frequency plots.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Bias</th>
<th>Noise</th>
<th>Gross Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired (in time and space) C</td>
<td>ΔC</td>
<td>ΔC - AC̄</td>
<td></td>
</tr>
<tr>
<td>Selectively paired C_m</td>
<td>ΔC_m</td>
<td>ΔC_m - AC̄_m</td>
<td></td>
</tr>
<tr>
<td>Unpaired in time, but paired in space</td>
<td>C_o &amp; C_p</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.2. Correlation. Spatial correlation can be demonstrated by use of transections and/or isopleth analyses in the horizontal (at the surface or aloft) or in the vertical (in a plane perpendicular or parallel to the mean wind) for point or urban sources. Temporal correlation can be illustrated by the use of time series plots, while space-time correlation can be demonstrated by the use of scatter diagrams. Specific data sets to be used in these analyses are shown in Table 4.

TABLE 4

Qualitative analyses techniques for correlation studies.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Temporal</th>
<th>Spatial</th>
<th>Space-Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired (in time and space) C</td>
<td>Time series of ΔC at a site</td>
<td>Isopleths and transections of ΔC at a time</td>
<td>Scatter diagrams of C_p or ΔC vs. C_o</td>
</tr>
<tr>
<td>Selectively paired C_m</td>
<td>Time series of ΔC_m for region</td>
<td>-</td>
<td>Scatter diagrams of (C_m)p or (ΔC)_m vs. (C_m)_o</td>
</tr>
</tbody>
</table>
Contingency tables, which function as quantitative scatter diagrams, were used by Roberts et al. (1970) and Zannetti and Switzer (1979) to evaluate air quality simulation models. The appropriate single valued performance measure to use in conjunction with such tables is the percent of incorrect or correct calculations (MacKay, 1981).

4. Error Analysis. Air quality simulation models have two sources of error according to Brier (1975), i.e., model errors and instrument error. In addition, improper use of statistical techniques can lead to erroneous estimates of the actual error.

4.1. Instrumentation. Instrument error leads to inaccuracies in measured concentrations. The magnitude of instrument error can be estimated by duplicate air quality monitoring instruments at individual sites (Brier, 1975). A quantitative analysis of instrument error was carried out by Anderson et al. (1977), who also supplied error bands for the maximum probable instrument error.

4.2. Model Error. Model error is composed of: (1) errors due to input data or from results from submodels used as input, and (2) errors due to deficiencies in the model itself.

Input error can be identified by the use of both case studies and sensitivity analysis, in which the effect of varying a single parameter is investigated. Such parameters include input parameters (e.g., pollutant emission strength, meteorological factors, and C0), forecast length, forecast period (e.g., time of day, day of week, or month of year), and location (horizontal or vertical position).

Input model errors associated with pollutant emission rates arise from uncertainties in the spatial distribution of sources, both in the horizontal and vertical. This last effect was investigated by Koch and Thayer (1970) who performed a sensitivity analysis by changing a uniform stack height to a distribution of stack heights. Another source of input model error is concerned with area source emission grid size. For example, Shieh et al. (1970) performed a sensitivity analysis which demonstrated that a small emission grid gave a more detailed concentration field than a large grid. Peak values of predicted concentration were also increased with the smaller grid.

The temporal variation of pollutant source strength is normally parameterized by a single meteorological variable, i.e., temperature. However, Shieh and Shir (1976) found systematic errors when they plotted observed and predicted SO2 values against air temperature. This parameter does not directly appear 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 accurate for extreme temperatures.
To formulate a better area source emission algorithm, it was reasoned that the 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 correction was applied to the area source emission formulation, systematic errors disappeared.

Similar results were found by Koch and Thayer (1971), who also carried out a complete sensitivity study involving other meteorological variables such as mixing depth and wind direction. The location of maximum predicted concentrations was extremely sensitive to small errors in this last parameter. A small error in input wind direction can result in a well predicted concentration pattern being shifted in location, resulting in a poor model evaluation. This is more of a problem with sharp gradients associated with narrow point source plumes, as opposed to the more uniform concentration fields associated with area source urban plumes. This is likewise more of a problem with primary pollutants, as opposed to secondary pollutants, for the same reason. Another problem with observations of point source plumes is the large number of zero values.

The probability of observing the maximum concentration is thus also lower for point source plumes and primary pollutants than for area source urban plumes and secondary pollutants. This is also true for relatively sparse observational networks associated with "real world" applications, as opposed to those of large scale research projects.

A final input parameter which can be a source of error is background concentration. Observed concentrations used by Mills and Stern (1975) were corrected for this effect before comparison to predictions from a large point source.

Sensitivity analysis must not be confused with "model tuning", in which constants within the model are adjusted to match observed and predicted concentrations to within a desired degree of accuracy (Bowen, 1980). Once this process is completed, the model formulation should be "frozen", and only data sets not used in the tuning process should be used in the evaluation process.

Output model errors arise due to deficiencies in the model itself in association with either parameterization of physical processes affecting concentration (such as diffusion parameters, linear decay and removal rates, and plume rise algorithms) or due to numerical approximations (e.g., artificial diffusion associated with modeling transport processes).

A theoretical mathematical analysis of instrument, input, and model errors was carried out by Brier (1975). However, he did not apply any of the suggested performance measures.

4.3. Statistical Errors. Statistical aspects of model evaluation processes can lead to erroneous estimates of model errors. Statistical
errors are related to inappropriate data and to improper use of statistical techniques.

In order to obtain appropriate data care must be taken in the design of air quality monitoring networks with respect to the number and placement of sites, as well as sampling frequency. In addition, during statistical analysis, effects of intercorrelations between sites must be accounted for.

Another such problem arises as pollutant observations are point values, while predicted values from many models are volume-averaged. Another aspect of this problem is that large differences exist between predicted rooftop and observed urban canyon concentrations (Johnson et al., 1970). They also discussed the need to relate these two concentrations via wind and traffic density data. Large (factor of two) variations in concentration were found by Johnson et al. (1971) within an urban canyon at the intersection of two streets due to the complex microscale circulation pattern within the canyon.

Problems arise, according to Bowne (1980), when statistics are estimated with more than one-third of the values missing. To reduce the influence of a few outliers on the variance, Brier (1975) suggested transforming the distribution of observed concentration to its logarithm or to itself raised to some power. Another approach is the use of robust (resistant) statistics such as the average absolute error instead of $\sigma$.

A basic averaging time of one hour should be used in all calculations (Fox, 1980). These values can then be used to compute the longer term three, eight, and twenty-four hour averages of the air quality standards.

Statistical problems arise when running averages are used to smooth data (Bowne, 1980), logarithmic plots are used (Bornstein, 1980), background concentration is removed from $C_o$ (Fox, 1980), and $C_o$ and/or $C_p$ are nondimensionalized by various combinations of $C_o$ and/or $C_p$ (Fox, 1980). However, nondimensionalization is frequently useful in presenting data, and various forms have been used by Hilst (1970), Roberts et al. (1970), Hayes (1979), and Rao et al. (1980).

Statistical problems also arise if zero predicted values are not eliminated before the cumulative frequency distribution is compared to that of the observed values (Bowne 1980). One method of eliminating zero values is by use of difference values.

Both Bowne (1980) and Brier (1975) discussed the need for symmetry (i.e. normality) in the distributions of $C_o$ and $C_p$, e.g., it is assumed in many statistical tests. While the distribution of $\Delta C$ values will be more normal than those of $C_o$ and $C_p$, the AMS Workshop
recommended using an exponential or logarithmic transformation of $C_0$ and/or $C_p$ to achieve normality. However, Brier (1975) suggested using robust statistics such as: (1) $|\Delta(\cdot)|$ instead of $\Delta(\cdot)$; (2) the median error instead of $\sigma$; (3) looking at the percentage of time a particular concentration is exceeded; and (4) looking at the high end of the frequency distribution of concentrations, as percentage error is more important there than at the low end of the distribution.

The problem of time correlations within data sets can be overcome using autocorrelation techniques to determine their magnitude (Fox, 1980). Once such trends are removed, data will have been converted to random variables, more easily analyzed by statistical techniques. Another technique is to use only some values, e.g., every n-th observation.

While $r^2$ is equal to the explained portion of the variance, correlation coefficients only indicate whether the "trend" is properly simulated. They do not provide quantitative estimates of the magnitude of the error, e.g., if the model always overpredicts concentrations by a constant factor, the correlation will be perfect. In addition, when observed data possess an exponential distribution (at a fixed location) or a log normal distribution (with respect to a moving coordinate system) logarithmic correlation coefficients should be computed (Shir and Shieh, 1976).

Nonlinear correlation coefficients could be used when appropriate and median, as opposed to mean, temporal and spatial correlation coefficients were reported by Duewer et al. (1978). In four of five cases, temporal correlation coefficients were higher than their corresponding spatial coefficients. Similar results were also found by Nappo (1974) in a comparison of nine multisource urban air quality simulation models. Only one of the models simulated mean spatial correlations better than mean temporal correlations. Finally, Brier (1975) has noted the necessity for the variance of the departures from the linear regression line to be constant.

Since, as discussed above, small errors in input wind directions cause well simulated concentration patterns displaced in space, spatial correlation coefficients are frequently poor. Thus MacKay (1981) cautions against using this technique and suggests instead the use of graphical techniques to demonstrate spatial correlation.

5. Conclusion. This paper has summarized various statistical techniques and performance measures used to evaluate predictions made by air quality simulation models. Various sources of error and practical problems associated with model evaluation were also discussed.

The suitability of the statistical testing of appropriate null hypotheses by evaluation studies is an area needing clarification. A statistical test investigates the hypothesis that observed and predicted
concentrations come from the same population. A test statistic is then calculated and compared with critical values, and the hypothesis is either rejected or accepted on the basis of this comparison. This approach was recommended by the AMS Workshop (Fox, 1980 and 1981).

An alternate argument is that the observed and estimated populations are assumed different and that small differences in their populations are statistically significant with large data bases. Thus, the appropriate question is whether or not these differences are significant in terms of a decision making process. It is therefore more informative to state confidence limits about the values of the performance measures. The AMS Workshop (Fox, 1981) recommended use of interval statements in evaluation studies and as an initial step in establishing future model performance standards.

A semi-qualitative analysis of the limitations of short term predictions due to the stochastic nature of concentration data was presented by Venkatram (1979). He proposed a method for estimating the difference between measured average concentrations and model predictions (which correspond to ensemble averages). The method involves computing the ratio between the observed maximum concentration during some time period and the predicted concentration. Details can be found in his paper, and it is hoped that future collaborations between air quality models and statisticians will help in determining the theoretical limiting accuracy of air quality simulation models.

Acknowledgement. This work was supported by SIAM Institute for Mathematics and Society (SIMS) and by EPA through its cooperative agreement with the AMS on air quality models.

REFERENCES


