Use of a Performance Based Approach to Determine Data Quality Needs
for the PM-Coarse (PMc) Standard

Executive Summary

Data quality objectives are qualitative and quantitative statements derived from the DQO Process that clarify study objectives, define the appropriate type of data, and specify the tolerable levels of potential decision errors that serve as the basis for establishing standards for the quality and quantity of data needed to support decisions.

Using some of the same techniques that were used to develop DQOs for fine particulate NAAQS ($PM_{2.5}$), the EPA developed a DQO software tool that provides decision makers with an understanding of the consequences of various input parameters, such as sampling frequency, data completeness, precision and bias and how these uncertainties affect the probability of making decision errors. Since both manual and continuos (automated) methods may be proposed for use in estimating the coarse particulate fraction, and the measurement uncertainties are unique to both methods, the DQO process can help weigh the benefits and disadvantages of these methods.

Preliminary data was collected from sites providing coarse particulate estimates from around the country as well as data from current multi-site performance evaluations conducted by the EPA National Environmental Research Laboratory. This data provided estimates of reasonable input parameters that were used to generate decision error performance curves. Preliminary decision error performance curves will be reviewed for effects of varying input parameters such as precision, bias, sampling frequency and completeness on both continuous and manual methods.

The DQO software provides user-friendly insights into the effects of uncertainty on decision making and identifies that the annual standard gray zones are most sensitive to population variability, sampling frequency, measurement bias, and completeness. The daily standard is sensitive to the variables listed above in addition to precision.

The goal of the paper is to provide details on the DQO approach taken and to elicit comments from CASAC as to the merits of this approach.

Discussion

DQOs are qualitative and quantitative statements derived from the DQO Process that clarify the monitoring objectives, define the appropriate type of data, and specify the tolerable levels of measurement errors for the monitoring program. By applying the DQO Process towards the development of a quality system, the EPA guards against committing resources to data collection efforts that do not support a defensible decision. The Office of Air Quality Planning and Standards (OAQPS) is contemplating the development of a particulate matter coarse (PMc) National Ambient Air Quality Standard (NAAQS). Since OAQPS developed a DQO in 1997 for $PM_{2.5}$, it was felt that an effort should be made to develop a DQO for PMc prior to any promulgation in order to provide decision makers with some idea of the ramifications of data uncertainty.

Decision makers need to feel confident that the data used to make environmental decisions are of adequate quality. The data used in these decisions are never error free and always carry some level of uncertainty. Because of these uncertainties, there is a possibility that decision errors can be made when measurements appear to provide an estimate above some action limit when the true estimate is below, or below an action limit when the true estimate is above. Therefore, decision makers need to understand and set limits on the data uncertainties that lead to these types of decision errors. The DQO process allows one to identify these data uncertainties, determine their effect on data quality and develop quality systems and network designs to reduce or maintain these uncertainties within acceptable levels. The intent of this paper is to describe the process used to identify data uncertainties, and by using this information, develop a DQO tool to help decision makers and those required to implement the monitoring program develop a quality system for PMc.

The DQO Performance Curve

OAQPS used performance curves to determine the effect of various types of uncertainties on decision error. Figure 1 is an example of a performance curve. The terms used in the figure are explained below:

Action limit - The action limit is the concentration or value that causes a decision maker to choose one of the alternative actions. A good example of action limits are the NAAQS standards where a concentration is identified and used to determine attainment or alternatively nonattainment of the NAAQS

Performance curves - Two performance curves have been generated based upon a number of input parameters of population and measurement uncertainties. The points along the curve are the true unknown concentration. The reason for the two curves is to represent measurement bias. The curve on the left side of the action limit represents the true concentration and the decision error relative to a positive 10% bias (as well as the other uncertainty values) while the curve on the right hand side of the action limit represents a true concentration and the decision error relative to a negative 10% bias (as well as the other uncertainty values).

Decision Error Limits - These limits are established by the decision makers and describes the decision makers' "comfort" with making a decision error, in the sense that a different decision would have been made if the decision maker had access to "perfect data" or absolute truth. The decision error limit in this example is 5%.

Gray Zone - The gray zone is the area between the performance curves where the decision errors are larger than the decision error limits; where the high cost or resources required to "tighten" the gray zone outweigh the consequences of choosing the wrong course of action.

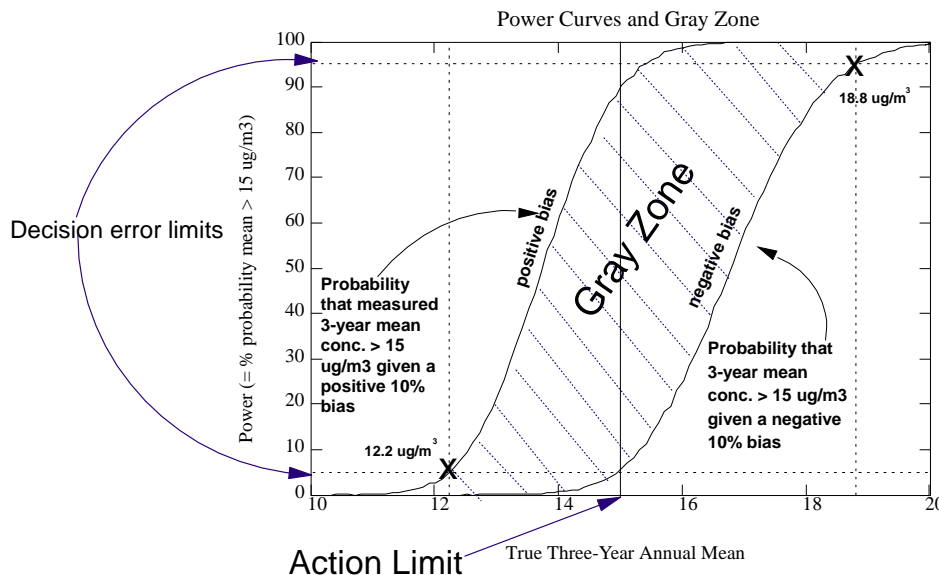**Power -** This is the probability of deciding that an observed design value exceeds the action limit.

Power Curves and Gray Zone

Figure 1 Example of a DQO Performance Curve

From Figure 1 the following statements could be made:

▸   If the true estimate is 18.8 ug/m^3 and if the measurement system has a negative bias of 10%, then 95% of the time the observed estimate will be above the 15 ug/m^3 action limit (correct decision) and 5% of the time the observed estimate will be less than 15 ug/m^3.

▸   If the true estimate is 12.2 ug/m^3 and the measurement system has a positive bias of 10%, then 5% of the observed estimates will be greater than 15 ug/m^3 and 95% will be less (correct decision).

▸   If bias of $\pm$ 10% is tolerable, any true estimate in the range of 12.2 to 18.8 ug/m^3 may have decision errors greater than 5%. As an example, for an estimate that truly is 17 ug/m^3 and the measurement system has a 10% negative bias, then 50% of the observed estimates will be declared to be less than the 15 ug/m^3 action limit .

The performance curve is a powerful tool for illustrating the effect uncertainties can have on the probability of making correct decisions. For example, larger biases widen the gray zone, while higher data completeness narrows the gray zone. Generally, the "steeper" the performance curves or the narrower the gray zone, the higher the probability of making correct decisions around the action limit. Thus, the performance curves can identify those uncertainties that have the greatest influence on decision errors, and help focus resources to minimize those uncertainties.

Sources of Uncertainty

Decision errors can be affected by the following variables related to four general categories: the method, the NAAQS, the sample population or the measurement uncertainty.

Uncertainty Related to the Method

There is a possibility that both integrated manual methods and continuous methods may be used to estimate PMc. One type of integrated method that is considered manual would require the use of two filter based sampling instruments; a PM_{10} instrument and a $PM_{2.5}$ instrument, where PMc would be estimated by subtracting the $PM_{2.5}$ estimate from the PM_{10} estimate. Using two instruments creates a potential for greater uncertainty, thus widening the gray zone. Automated PMc instruments are available and have the advantage of continuous sampling but these instruments are still under development and display some bias in certain geographic areas. Historically, for each ambient air criteria pollutant, one method type is designated as a federal reference method (FRM). The manual methods for PM_{10} and $PM_{2.5}$ are currently designated as FRMs and may need to be used in PMc to provide an estimate of bias for the continuous methods.

Uncertainty Related to the NAAQS

▸   **Level of standard** - The level of the standard refers to the concentration where the action limit is set. For example, if an action limit is set at a concentration close to the sensitivity of the method, one would expect more potential for decision error. The information on the potential concentration ranges of the two standards is included in the Draft EPA Staff Paper: *Review of the National Ambient Air Quality Standards for Particulate Matter*. Since the standard has not been promulgated, OAQPS used the max/min of the annual and daily standard identified in the Staff Paper (see Table 1)

▸   **Form of the standard** - If one uses an annual average versus the highest concentration in a year, there would be more potential for decision error with the single high concentration value. Current thinking on the PMc is to propose two standards similar to the current $PM_{2.5}$ standard: a three year annual average value (annual average) and a 3-year percentile of a 24 hour average value (daily standard). OAQPS developed DQO scenarios for both forms.

▸   **Percentile for daily standard** - different percentiles of the daily standard could affect decision error. OAQPS looked at 98, 95 and 90 percentiles of a 3-year 24 hour average but did not notice significant differences in the DQO gray zone, so a 98th percentile was used.

Uncertainty Related to Sample Population

Values related to sample population were developed through a review of PM_{10} and $PM_{2.5}$ data available in AQS. Values for each attribute were selected at a conservative but realistic level, meaning that 90-95% of the sites had values less than (which would narrow the gray zones) the ones chosen for input to the DQO performance curves. Population uncertainty inputs, once selected, are not changed when running DQO performance curves scenarios.

▸   **Seasonality ratio** - the ratio of the highest concentration to the lowest concentration within a particular time period. A ratio of 7 for PMc was used.

▸   **Population variability** - measures the random, day-to-day movement of the true concentration about the average sine curve. 60% for PMc was used.

▸   **Autocorrelation** - a measurement of the estimated similarity on successive days. Since there is a possibility that PMc can be measured on a 1 in 6 day sampling frequency, an autocorrelation of 0 was used.   If continuous instruments are used, everyday sampling will be viable and some autocorrelation may be incorporated into the DQO.

More details on estimating the population parameters can be found in Appendix A. Based upon the review of PMc data in AQS (see Appendix A) and the data from the NERL Intercomparison study, a set of input parameters to generate the performance curves were selected. Table 1 provides the population parameter estimates derived from the AQS data. For the DQO Tool, the second column (Sel.) identifies the parameter values used to generate the performance curves. As with the $PM_{2.5}$ DQO, the parameters chosen where conservative; producing a "wider gray" zone, but within realistic values of the data.

Table 1 Estimated particulate matter population parameters.

Quantile	Sel.	10	20	30	40	50	60	70	80	90	97.5
PM2.5 Ratio	5.3	1.46	1.63	1.77	2.02	2.14	2.28	2.58	3.03	4.01	5.72
PMc Ratio	14	1.68	2.05	2.32	3.24	3.82	4.42	5.54	8.01	14.3	52.52
PM2.5 CV	0.8	0.35	0.41	0.45	0.51	0.53	0.56	0.6	0.64	0.69	0.8
PMc CV	1	0.4	0.49	0.56	0.66	0.71	0.76	0.84	0.93	1.08	1.39
PM2.5 autocorrelation	0	0	0.06	0.25	0.42	0.45	0.48	0.51	0.59	0.68	0.94
PMc Autocorrelation	0	0	0.13	0.19	0.28	0.31	0.44	0.48	0.51	0.64	0.81
PMc to PM2.5 Ratio	2.25	0.28	0.37	0.46	0.72	0.87	1.04	1.29	1.6	2.22	3.29
Correlation	0	-0.23	-0.05	0.06	0.19	0.25	0.31	0.39	0.46	0.56	0.69

Uncertainty Related to Measurement System

▸   **Sampling frequency** - The DQO tool used both 1 in 6 day and every day sampling frequency to accommodate both manual and continuous methods.
▸   **Completeness** - 75% was used since it is currently allowed in CFR for particulate matter.
▸   **Measurement bias** - 10% bias was used as this appears reasonable for $PM_{2.5}$ and would probably remain reasonable. Table 2 provides an estimate of bias of the various direct and indirect methods used in the NERL PMC method intercomparison study. Two estimates of bias are provided; mean bias where positive and negative bias can cancel and absolute bias (abs) where the absolute value of each individual bias estimate is used and the mean taken from those individual values. The absolute value bias has been proposed as the new bias statistic for the gaseous pollutants in CFR and would be proposed to use as the estimate for PMc. The statistics used in Table 2 are described in Appendix B. Since the mean absolute bias statistic uses absolute values, it does not have a tendency (negative or positive) associated with it. A sign will be designated by rank ordering the relative percent differences (with signs) for site values. Calculate the 25th and 75th percentiles. The absolute bias would be flagged as positive (+) if both the 25th and 75th percentiles where positive and negative (-) if both are negative. The mean absolute bias would not be flagged with a sign if the percentiles were different signs. The mean absolute bias estimates in Table 2 reflect this process Of genuine concern for the personnel trying to develop the PMc quality system is the instrument or standard for use in determining bias for this network.

Table 2- PMc Bias Estimate with FRM_1 as Truth

	All sites PMc		Phoenix PMc		Gary PMc		Riverside PMc	
	mean bias	mean abs bias	mean bias	mean abs bias	mean bias	mean abs bias	mean bias	mean abs bias
APS	-45.0%	-52.3%	-45.4%	-45.4%	-45.8%	-64.2%	-43.6%	-43.6%
APS_1	-42.4%	-54.2%	-44.1%	-44.1%	-37.9%	-69.1%	-46.3%	-46.3%
APS_2	-47.8%	-50.1%	-47.2%	-47.2%	-53.7%	-59.2%	-40.9%	-40.9%
Dichot	-10.4%	-11.5%	-18.8%	-18.8%	-9.4%	-12.0%	-6.5%	-6.8%
Dichot_1	-12.1%	-12.5%	-18.7%	-18.7%	-11.1%	-12.2%	-6.2%	-6.6%
Dichot_2	-8.2%	-9.8%	-15.1%	-15.1%	-9.3%	-12.3%	-6.2%	-6.5%
Dichot_3	-10.8%	-11.9%	-19.8%	-19.8%	-8.8%	-11.5%	-7.2%	-7.4%
FRM	-1.0%	5.0%	1.2%	4.0%	0.2%	6.0%	-4.1%	-4.8%
FRM_2	-4.7%	-5.6%	-2.5%	-3.6%	-4.1%	-5.9%	-6.5%	-6.6%
FRM_3	2.1%	4.4%	3.7%	+4.2%	4.4%	+6.0%	-1.7%	3.1%
TEOM	-16.9%	-22.8%	7.0%	10.5%	-31.0%	-31.1%	-26.2%	-26.4%
TEOM_1	-17.8%	-25.8%	10.9%	+13.8%	-32.8%	-32.8%	-30.0%	-30.0%
TEOM_2	-20.8%	-23.2%	0.0%	6.8%	-32.7%	-32.7%	-30.6%	-30.6%
TEOM_3	-12.1%	19.3%	10.1%	+11.0%	-27.7%	-27.7%	-18.0%	-18.7%
Tisch	0.4%	9.9%	5.6%	8.4%	-8.9%	-13.8%	5.1%	7.2%
Tisch_1	-1.8%	10.9%	3.1%	+4.7%	-15.6%	-19.2%	7.8%	+8.4%
Tisch_2	2.3%	10.3%	14.1%	16.6%	-5.9%	10.2%	-0.7%	4.4%
Tisch_3	0.8%	8.3%	-0.3%	3.9%	-5.3%	12.1%	8.3%	+8.6%

Table 3- Precision estimates from NERL Intercomparison Study

monitor	number of complete runs	precision
all_FRM	76	4.1%
all_R_P_Dichot	47	3.8%
all_Tisch	90	10.1%
all_TEOM	84	6.2%
all_APS	70	19.5%

‣ **Measurement precision** - 10% precision was used as this appears reasonable for $PM_{2.5}$ and would probably remain reasonable for either a manual or continuous method. More information on this uncertainty is being assessed. Table 3 provides estimates of precision from the NERL Intercomparison study. The statistics used in Table 3 are described in Appendix B.

The DQO Software Tools

The DQO tools use performance curves which allows one to model $PM_{10-2.5}$ data based on the fixed population uncertainty assumptions. Then, the performance curves are changed based on the inclusion of measurement uncertainty input parameters of sampling frequency, precision, bias and completeness. The goal is to keep the gray zone as narrow and the performance curves as steep as possible. Two DQO software tools were developed: one, the direct measurement tool, can be used for continuous instruments or manual instruments that provide PMc material on a single filter; and a second tool, the integrated tool, when the method requiring a PM_{10} instrument and a $PM_{2.5}$ instrument, is used. The DQO tools allow one to generate multiple performance curve on the same graph by altering the measurement uncertainty values. By altering these uncertainty values, one can determine which uncertainty has the most effect on data quality. Figure 2 provides an example graph derived from the direct DQO tool where only sampling frequency was altered from 1 in 6 day to everyday.
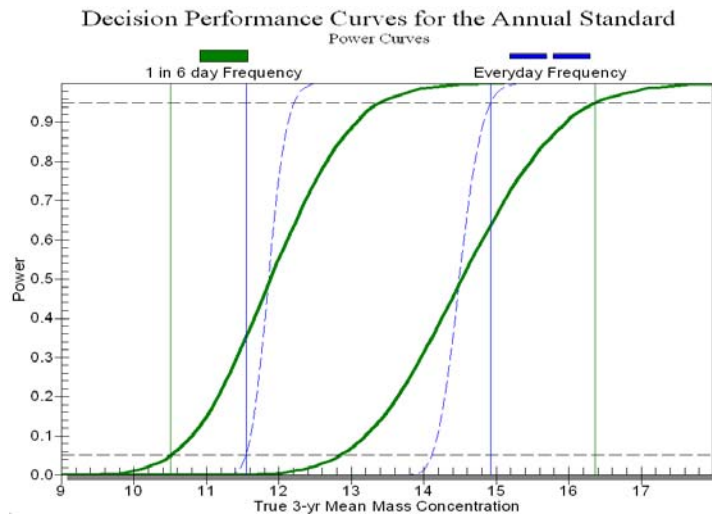
Decision Performance Curves for the Annual Standard
Power Curves

Figure 2 Direct DQO Tool Example

Performance Curve Results

Table 4 provides gray zones for PMc at the NAAQS levels mentioned in the Draft Staff Paper. This table provides an example of the changes in the width of the gray zone in relation to sampling frequency and sampling method. For PMc the gray zone estimates for columns 3 (1-6 day integrated) and 4 (every day integrated) were developed from the DQO software where two instruments (a PM_{10} and a $PM_{2.5}$) are used to derive a PMc concentration. The 5th column, identified as "Direct" is the gray zone derived either from a continuous instrument or an instrument collecting a coarse sample on one filter. The reason for the larger gray zones from the integrated method even when every day sampling occurs are related to the additive errors of two methods in order to derive a concentration.

Table 4 - PM $_{10-2.5}$ Gray Zone Performance Curve Values Relative to Sampling Frequency and Method

Standard	NAAQS ug/m³	1-6 Day (Integrated)	Every Day (Integrated)	Every Day (Direct)
Annual	13	9.9 - 18.0	10.6 - 16.6	11.6 - 14.8
	30	22.9 - 41.6	24.4 - 38.3	26.7 - 34.1
Daily 98%	30	17.6 - 41.9	24.2 - 38.7	25.3 - 35.7
	75	43.1 - 102.9	59.7 - 95.7	62.6 - 88.5

Conclusions

The DQO software provides user-friendly insights into the effects of uncertainty on decision making and identifies that the annual standard gray zones are most sensitive to population variability, sampling frequency, measurement bias, and completeness. The daily standard is sensitive to the variables listed above in addition to precision. Results from the DQO work are preliminary. The information on the form and the level of the standard are draft proposals and are used only to provide an example of the DQO software's capability. The population and measurement uncertainty parameters have not been agreed upon and may change, thus changing the gray zones. The EPA National Environmental Research Laboratory is currently conducting intercomparisons on a number of the PMc manual and continuous instruments. This information will be used to check the population and measurement uncertainty assumptions in order to revise the software as needed and to provide more accurate assessments of the probability for decision errors**.**

Appendix 3A

Technical Report on Estimating Parameters
For the PMc DQO Tool

May 15, 2003

TECHNICAL REPORT

on

ESTIMATING PARAMETERS FOR THE PM$_{coarse}$ DQO TOOL

Contract No. 68-D-02-061
Work Assignment 1-01

for

Vickie Presnell, Project Officer
Shelly Eberly, Work Assignment Manager

Emissions, Monitoring, and Analysis Division
Office of Air Quality Planning and Standards
U.S. ENVIRONMENTAL PROTECTION AGENCY
Research Triangle Park, North Carolina 27711

Prepared by

Basil W. Coutant and Christopher H. Holloman

BATTELLE
505 King Avenue
Columbus, Ohio 43201-2693

TABLE OF CONTENTS

List of Tables

List of Figures

EXECUTIVE SUMMARY

Data Quality Objectives (DQOs) are being developed for PM_{coarse}. To aid in this development, a simulation model has been developed and has been implemented in a software tool as was done for $PM_{2.5}$ (U.S. EPA, 2002). This report describes the simulation modeling process and the parameters that control the process. The simulation model assumes that PM_{coarse} will be measured using a difference method, namely PM_{10} - $PM_{2.5}$, rather than a direct measurement. The parameters in the model describe the ambient behavior of $PM_{2.5}$, PM_{10}, and PM_{coarse}. Users must estimate these parameters from ambient data to determine the relevant ranges to be explored while developing the DQOs.

The parameters describing the ambient conditions are as follows: the degree of seasonality of ambient $PM_{2.5}$ and PM_{coarse} over the year, the day-to-day variability of ambient $PM_{2.5}$ and PM_{coarse}, the ratio of the mean level of ambient PM_{coarse} to the mean level of ambient $PM_{2.5}$, the difference between the times of year when ambient $PM_{2.5}$ and ambient PM_{coarse} peak, and the correlation between ambient $PM_{2.5}$ levels and ambient PM_{coarse} levels. Using data from 622 sites in EPA's Air Quality System (AQS) database, these parameters have been estimated at the site level to find the ranges that are likely to be encountered across the nation. Table ES-1 summarizes the findings for the 622 sites. The information contained in Table ES-1 is intended to guide users in exploring values relevant for their own development DQO. Detailed descriptions of the parameters in the table may be found in Section 3.0.

Table ES-1.   Estimated particulate matter parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
$PM_{2.5}$ ratio	1.46	1.63	1.77	1.88	2.02	2.14	2.28	2.58	3.03	4.01	5.72
PM_{coarse} ratio	1.68	2.05	2.32	2.73	3.24	3.82	4.42	5.54	8.01	14.34	52.52
$PM_{2.5}$ CV	0.35	0.41	0.45	0.48	0.51	0.53	0.56	0.6	0.64	0.69	0.8
PM_{coarse} CV	0.4	0.49	0.56	0.61	0.66	0.71	0.76	0.84	0.93	1.08	1.39
$PM_{2.5}$ autocorrelation	0	0.06	0.25	0.35	0.42	0.45	0.48	0.51	0.59	0.68	0.94
PM_{coarse} autocorrelation	0	0.13	0.19	0.22	0.28	0.31	0.44	0.48	0.51	0.64	0.81
k	0.28	0.37	0.46	0.56	0.72	0.87	1.04	1.29	1.6	2.22	3.29
correlation	-0.23	-0.05	0.06	0.12	0.19	0.25	0.31	0.39	0.46	0.56	0.69

Note that the PM_{10} concentrations used to develop Table ES-1 are in "standard conditions," meaning that the volume used in the calculation of the concentration was adjusted to the volume corresponding to 1 ATM and 25 C. The tool generally allows a larger range of values than is indicated by Table ES-1, in part to allow for changes to the reporting units.

Technical Report

on

Estimating Parameters for the PM$_{COARSE}$ DQO Tool

1.0 INTRODUCTION

The EPA guidelines for the Data Quality Objectives (DQO) Process (U.S. EPA, 2000) specify a seven-step procedure for creating DQOs. For the remainder of this document (QA/G-4), it is assumed that the reader is familiar with the procedure outlined in QA/G-4 and with using decision performance curves to evaluate the utility of decision rules. The purpose of the procedure outlined in QA/G-4 is to create DQOs that will assure the collection of data that are not only relevant to questions of interest, but also of high enough quality to ensure the ability to answer those questions. In the current context, users are interested in answering several questions including, but not limited to:

- What action levels (the National Ambient Air Quality Standards, NAAQS, or local standards) one should specify in order to limit ambient PM$_{coarse}$ levels to a suitable level?
- What levels of sampling frequency, data completeness, measurement bias, and vendor measurement error are needed to detect the concentrations efficiently?

To help users answer these questions, a DQO tool has been developed to quantify the uncertainty associated with policy decisions. Use of this tool requires users to perform five steps:

1. Obtain particulate matter data. (Historical data available in AQS is sufficient.)
2. Estimate parameters that describe important characteristics of these data to be input into the DQO tool.
3. Determine other values relevant to the process, such as acceptable bias in samplers, to be input into the DQO tool.

4.   Enter the values from Steps 2 and 3 into the DQO tool and initiate the calculations.

5.   Interpret the decision performance curves the DQO tool produces.

In this document, issues related to each of these steps are addressed. However, the main focus is on the calculations necessary to complete Step 2 and the research necessary to complete Step 3. In addition, this document gives information on the inner workings of the DQO tool in order to allow better understanding of the decision performance curves it produces.

Step 1. Obtaining Particulate Matter Data

Before obtaining PM data for development of DQOs, users need to give careful consideration to the boundaries of the study region of interest. After defining a region of interest, relevant data may be downloaded from the EPA's AQS database (see http://www.epa.gov/ttn/airs/airsaqs/index.htm).

Steps 2 and 3. Estimating Parameters and Determining Other Important Values

Section 3.0 of this document explains methods for determining inputs to be used by the DQO tool to produce useful decision performance curves. The relevant inputs are simple to specify. These inputs fall into three categories: those that describe inherent properties of ambient particulate matter, those that describe properties of the sampler, and those that describe the NAAQS and the quality of the decision. They include:

Properties of ambient particulate matter:

* The degree of seasonality of ambient $PM_{2.5}$ and PM_{coarse} over the year.
* The day-to-day variability of ambient $PM_{2.5}$ and PM_{coarse}.
* The ratio of the mean level of ambient PM_{coarse} to the mean level of ambient $PM_{2.5}$.
* The difference between the times of year when ambient $PM_{2.5}$ and ambient PM_{coarse} peak.

- The correlation between ambient $PM_{2.5}$ levels and ambient PM_{coarse} levels.

Properties of the sampler:

- The amount of random measurement error at the samplers.
- The amount of bias introduced by the samplers.
- The quarterly completeness of $PM_{2.5}$ and PM_{10} readings.
- The number of days between adjacent readings.

Quantities relevant to the NAAQS and the quality of the decision:

- Acceptable Type I and Type II error (defined in Section 3.7).
- The percentile for PM_{coarse} daily standard.
- The daily standard.
- The annual standard.

Obtaining information from the DQO tool requires estimates for the first group of parameters. Guidance for estimating these values is provided in Section 3.0 of this document. Guidance for estimating the quantities that describe properties of the sampler may be obtained from national QA reports for the PM_{10} and $PM_{2.5}$ network or operator experience. Guidance for specifying values for Type I error and Type II error can be obtained from EPA QA/G-4. This document specifies a guideline of 1 percent as the starting point for determining acceptable levels of Type I and Type II error. The remaining quantities are specified at the discretion of the user and/or the yet to be determined NAAQS for PM_{coarse}.

Rather than using the data to calculate a single set of parameter estimates to be input into the DQO tool, users should attempt to determine a range of plausible values for the parameters. For instance, if the chosen region of interest is the entire nation, the parameters should be estimated for individual sites across the nation and consideration should be given to the range of values that parameters take at different sites. This strategy of calculating parameter estimates separately for each site within the region of interest can be applied whenever multiple sites exist

in the region of interest. In fact, the region of interest should be expanded to include several sites. At the very least, users should perform the analysis multiple times using not only the estimated values, but also other parameter values close to the estimated values that could be considered worse (e.g., less frequent data collection or larger measurement bias).

Steps 4 and 5: Running the DQO Tool and Interpreting the Results

The DQO tool is simple to operate – users enter information obtained in Steps 2 and 3 into the graphical interface and click a radio button to initiate the calculations. (Details of the operation of the DQO tool may be found in the "User's Guide for DQO Companion for PMcoarse.") While not strictly necessary, an understanding of the methods used to calculate the decision performance curves may aid the user with their interpretation. Section 2.0 and Appendix A explain the inner workings of the DQO tool.

The remainder of this document gives technical details on the DQO tool and on estimation techniques users can employ to determine inputs to the tool. Users who are familiar with the DQO tool for $PM_{2.5}$ (U.S. EPA, 2002) and are uninterested in the details of simulation may skip to Section 3.0. Users who are unfamiliar with decision performance curves and their interpretation may find that understanding the methods used by the DQO tool help with understanding the output the tool produces. Those users are encouraged to read Section 2.0. Users interested in the technical details of the statistical model underlying the simulations described in Section 2.0 are referred to Appendix A.

2.0 SIMULATION

This section describes the methods the DQO tool uses to calculate decision performance curves after being given inputs from the user. All of the calculations and estimation procedures described in this section are performed by the DQO tool and require no user interaction.

The DQO tool uses inputs that describe the physical state of nature and the properties of the samplers to simulate different scenarios of observed readings. In this way, simulation acts as a bridge between decision maker inputs and decision maker/data user requirements for data quality. The DQO tool bridges the gap between the user inputs and the data quality requirements in three steps:

Step 1.   The DQO tool uses the inputted values to simulate a physical process that mimics the true behavior of particulate matter.

Step 2.   The DQO tool adjusts the simulated particulate matter values from Step 1 to account for bias, measurement error, and missing data in order to mimic data collection.

Step 3.   The DQO tool calculates decision performance curves and gray zones corresponding to the simulated data obtained in Step 2.

To make this process work, the simulation model needs to mimic the major properties of the physical process. The values of these properties (estimated using the techniques described in Section 3.0) are input into the simulator by the user. These properties include:

- **The degree of seasonality of ambient $PM_{2.5}$ and PM_{coarse} over the year**. The simulator assumes that typical $PM_{2.5}$, PM_{coarse}, and PM_{10} levels follow sinusoidal patterns over the course of a year. The ratio of the peak of the seasonal level to the trough of the seasonal level is specified by the user.
- **The day-to-day variability of ambient $PM_{2.5}$ and PM_{coarse}**. Both of these types of particulate matter exhibit random variability around their seasonal mean sinusoidal behavior.
- **The ratio of the annual mean level of ambient PM_{coarse} to the annual mean level of ambient $PM_{2.5}$.**
- **The difference in months between the times of year when ambient $PM_{2.5}$ and ambient PM_{coarse} peak**.

- **The correlation between ambient PM$_{2.5}$ and ambient PM$_{\text{coarse}}$.** Levels of PM$_{2.5}$ and PM$_{\text{coarse}}$ tend to vary together in many areas. The degree to which their ambient levels vary together is an important characteristic to simulate.
- **The amount of random measurement error at the samplers**.
- **The amount of bias introduced by the samplers**.

Additionally, there are known decision-maker constraints to the process that affect the output, including:

- **The quarterly completeness of PM$_{2.5}$ and PM$_{10}$ readings**. Current EPA guidelines for data quality stipulate that no more than 25 percent of data readings may be missing during each quarter when calculating summary statistics for PM$_{2.5}$ (40 CFR 58). The DQO tool allows different proportions of missing data in order to assess its impact on data quality.
- **The number of days between adjacent readings**. Particulate matter readings are often not available on a daily basis, but instead on a "1 in m days" basis. The DQO tool accounts for this type of data collection.

Once satisfactory values for these variables have been determined, the user must also specify a percentile and daily standard to be monitored as well as acceptable levels of Type I and Type II error (explained in Sections 3.7 through 3.9).

The DQO tool calculates the probability of making NAAQS-like non-attainment decisions for both an annual and a daily standard as a function of the true 3-year values. The DQO tool calculates these probabilities of interest via simulation. In order to create the decision performance curves, the simulator first creates 5,000 instances of PM$_{2.5}$ and PM$_{10}$ data. Each of these 5,000 instances contains data covering a 3-year span. These 5,000 sets of three years' worth of data are depicted within the rectangle in the upper left corner of Figure 2-1.

First, consider only one of these 5,000 instances (the first instance is singled out in the figure). Once the data for this instance are generated using the user specified parameters, both

the $PM_{2.5}$ and PM_{10} series are scaled (multiplied by a constant) so that a value of 1 unit corresponds to the percentile for the daily standard for $PM_{coarse} = PM_{10} - PM_{2.5}$ for the simulated truth. The figure depicts this scaling when the percentile of interest is the 98th percentile.
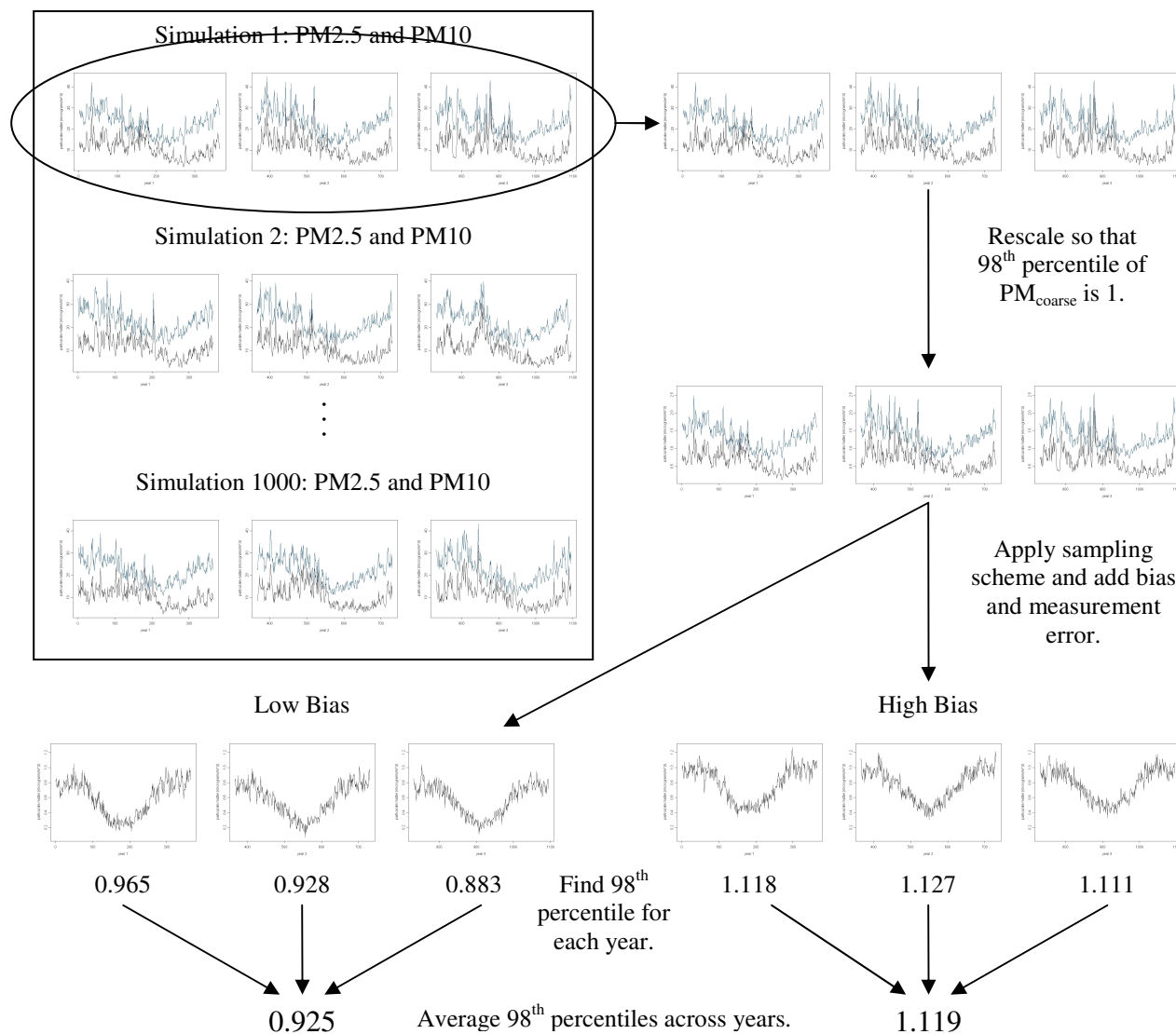
Figure 2-1. Scheme for simulation of data and calculation of percentiles.

Next, for each quarter of each year the completeness and "1 in m days" sampling restrictions are applied, and measurement error and bias are introduced into the observations to

account for data collection conditions. Bias is applied in a "best case" and "worst case" scenario – in one instance, $PM_{2.5}$ readings are biased up while PM_{10} readings are biased down resulting in underestimation of PM_{coarse} levels; in the other instance, $PM_{2.5}$ readings are biased down while PM_{10} readings are biased up resulting in overestimation of PM_{coarse} levels. The result is that the three years of simulated data are turned into two sets of three years of data: three years of the high bias case and three years of the low bias case.

For now, consider only one of these bias cases (although both cases are depicted in the figure). In this bias case, the DQO tool uses the simulated data to calculate the level of PM_{coarse} that corresponds to the percentile for the daily standard for each of the three years of data. These three levels (one for each year) will be averaged to get a mean annual percentile level that corresponds to the percentile of interest. This process is repeated for the other bias case, and the entire process is repeated 5,000 times to get 10,000 estimates of the PM_{coarse} value that corresponds to the percentile of interest. Those 10,000 estimates will consist of 5,000 estimates corresponding to the low bias case and 5,000 estimates corresponding to the high bias case. Details of the mathematical models used for the simulation described are found in Appendix A.

Once the DQO tool has calculated these numbers, it can calculate the probability of an observed average annual percentile exceeding the daily standard. Suppose that we are interested in the probability that the observed 98th percentile exceeds a 35 $\mu g/m^3$ standard. Suppose also that the true 98th percentile is 20 $\mu g/m^3$. Recall that the PM_{coarse} levels simulated are scaled so that the true PM_{coarse} level corresponding to the 98th percentile is 1 unit in the unbiased case. Since we are interested in the case where the true 98th percentile is 20 $\mu g/m^3$, we can multiply all of the 10,000 estimates we obtained by 20. Then, to find out the probability that the 98th percentile exceeds 35 $\mu g/m^3$, the DQO tool calculates the proportion of the 5,000 simulations in each of the high and low bias cases that exceed 35 $\mu g/m^3$. These proportions are the probabilities that are plotted as decision performance curves when this evaluation is performed for many different possible true 98th percentile values. Two decision performance curves are plotted: one for the high bias case and one for the low bias case. Figure 2-2 shows two of the decision performance curves described.
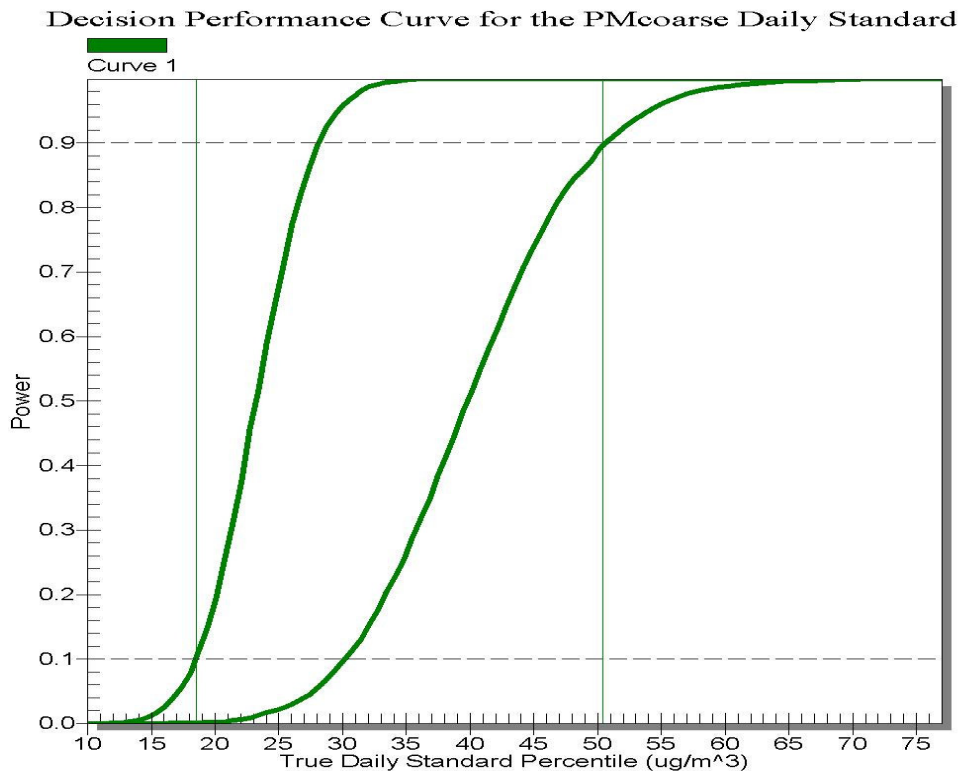
Figure 2-2. **Example decision performance curves.**

Once the decision performance curves have been plotted, horizontal lines are added to the plots at the levels of acceptable Type I and Type II errors as shown in Figure 2-2 (the lines are added at 10 percent Type I error and 10 percent Type II error in Figure 2-2). From the intersection of these lines with the decision performance curves, the DQO tool calculates the gray zone. The gray zone is the range of true values over which Type I and Type II errors will be unacceptably large. In Figure 2, the gray zone is from 18.5 to 50.5. The DQO tool can help determine what data collection quality assurance measures are necessary to reduce the gray zone to a desired size.

Note that, so far, we have only addressed how the simulator calculates the probability of an observed percentile of daily readings exceeding a standard. The simulator also creates decision performance curves related to the attainment of an annual standard. To calculate the

9

probability that the annual mean exceeds the standard, the $PM_{2.5}$ and PM_{10} simulations are rescaled so that instead of the fixed percentile (the 98th percentile, in this example) of the PM_{coarse} series being 1 unit, the mean of the true PM_{coarse} series is 1 unit. Then, the rest of the calculations are performed in the same way as before – bias and measurement error are added, and decision performance curves are calculated using the 5,000 replications from each of the high and low bias cases. The result is a second set of decision performance curves that show the probability of the observed mean of PM_{coarse} exceeding the annual standard for several true values of the mean of PM_{coarse}.

3.0 PARAMETER ESTIMATION TECHNIQUES

To use the PM_{coarse} DQO tool, users must estimate the parameters that describe ambient particulate matter properties for the specific geographic region of interest. These estimates are then input into the PM_{coarse} DQO tool to create relevant decision performance curves. Sections 3.1 through 3.6 explain the parameters that describe true ambient particulate matter properties. These sections describe how to estimate the parameters from data collected from the region of interest and also describe what these parameters measure. Sections 3.7 through 3.13 explain the parameters describing properties of the sampler and the parameters describing decision rules and acceptable error rates.

PM$_{2.5}$ and PM$_{10}$ data for most areas can be obtained from the EPA's AQS database (see http://www.epa.gov/ttn/airs/airsaqs/index.htm). Both sets of data are needed for the region of interest. PM_{coarse} values are obtained by subtracting $PM_{2.5}$ from PM_{10} on a day-by-day basis. Note that calculating PM_{coarse} values in this way can result in some negative PM_{coarse} values if PM_{10} readings fall below $PM_{2.5}$ readings at some time points (this situation is possible with biased samplers). It is assumed that whenever this occurs, the PM_{coarse} value should be set to zero for that day. Once the data have been obtained and PM_{coarse} values have been calculated, calculation of the DQO parameters to be input into the DQO tool can be performed as outlined in the following sections.

In addition to the parameter descriptions and explanations of how to estimate them, the following sections contain summaries of values found for each parameter across several sites in the United States. These summaries were calculated using data downloaded from the AIRS/AQS database. All PM_{10} measurements used in these calculations are in standard units, meaning that the volume used in calculating the concentrations was adjusted to 1 ATM and 25 C. The parameter value summaries were created for use in developing national level DQOs and are also intended as a guide for users estimating their own parameters to be input into the model.

3.1   Seasonality Ratio

The ratio parameter is a measure of the degree of seasonality in the data. It is the ratio of the high point to the low point on the sine curve that describes the average behavior of PM. This ratio must be estimated separately for the $PM_{2.5}$ and PM_{coarse} series. With at least one year of data, the ratio can be estimated by calculating the means for each month and dividing the highest value by the lowest value. With more than one year of data, each month of data is averaged, even though the individual values may come from different years. The ratio of the maximum monthly average to the minimum monthly average is an estimate of the true ratio parameter. Table 3-1 shows several quantiles of the estimated ratio parameter for $PM_{2.5}$ and PM_{coarse} across 502 sites in the United States. For ease of use, the values in this table are repeated in Section 4.0.

Table 3-1.   Quantiles of the estimated seasonality ratio parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
$PM_{2.5}$ ratio	1.46	1.63	1.77	1.88	2.02	2.14	2.28	2.58	3.03	4.01	5.72
PM_{coarse} ratio	1.68	2.05	2.32	2.73	3.24	3.82	4.42	5.54	8.01	14.34	52.52

3.2   Population Coefficient of Variation (CV)

This parameter measures the amount of random, day-to-day movement of the true concentration around the average sine curve. Again, the population coefficient of variation (CV) parameter is estimated separately for the $PM_{2.5}$ and PM_{coarse} series. This parameter is slightly

11

more difficult to estimate than the ratio parameter. First, a sequence of the natural logarithms of the concentrations is generated from measurements that were taken <u>every 6th day</u> (deleting if needed). Next, a new sequence of numbers is generated equal to the differences of successive pairs in the sequence of the logs. Every other term in this sequence is removed. The standard deviation of this set of numbers is calculated and referred to as $S6$. An estimate for the population CV is $\sqrt{\left(\exp\left(S6^2/2\right)-1\right)}$. Table 3-2 shows several quantiles of the estimated CV parameter for both $PM_{2.5}$ and PM_{coarse} across several sites in the United States. For ease of use, the values in this table are repeated in Section 4.0.

Table 3-2.   Quantiles of the estimated population CV parameter

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
$PM_{2.5}$ CV	0.35	0.41	0.45	0.48	0.51	0.53	0.56	0.6	0.64	0.69	0.8
PM_{coarse} CV	0.4	0.49	0.56	0.61	0.66	0.71	0.76	0.84	0.93	1.08	1.39

3.3   Autocorrelation

Another parameter describing the natural variability of the true concentrations is autocorrelation. Like the preceding variables, the autocorrelation is estimated separately for the $PM_{2.5}$ and PM_{coarse} series. This is a measurement of the similarity between successive days. Estimating autocorrelation is harder than estimating the population CV. Without daily measurements, the value of 0 should be used. Realistically, 0 is the most conservative case and can always be used. Assuming daily measurements are available, let S6 be computed as in Section 3.2 (the standard deviation computed from differences of the logs from <u>every 6th day</u> measurements). Let $S1$ be computed in the same manner using differences of logs from daily measurements. If $S6 > S1$, then there is some autocorrelation, and it is estimated with $\left(S6^2 - S1^2\right)/S6^2$. This formula tends to slightly overestimate the true amount of autocorrelation present in the data. Since it is better to underestimate this parameter (to make the results more conservative), the user may want to multiply the estimate by 0.85. Table 3-3 shows several quantiles of the estimated autocorrelation for both $PM_{2.5}$ and PM_{coarse} across several sites in the

United States with daily measurements. The estimates reported in Table 3-3 below were not
multiplied by 0.85. For ease of use, the values in this table are repeated in Section 4.0.

Table 3-3.   Quantiles of the estimated autocorrelation parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
$PM_{2.5}$ Autocorrelation	0	0.06	0.25	0.35	0.42	0.45	0.48	0.51	0.59	0.68	0.94
PM_{coarse} Autocorrelation	0	0.13	0.19	0.22	0.28	0.31	0.44	0.48	0.51	0.64	0.81

3.4 PM_{coarse} to $PM_{2.5}$ Ratio

This parameter is used for scaling within the simulation model. Its estimation is very
simple. First, let M1 be the average of all of the PM_{coarse} values over the full time period
available. Let M2 be the same average for the $PM_{2.5}$ data. Then, the ratio is estimated with
M1/M2. Table 3-4 shows several quantiles of the estimated value of this ratio across several
sites in the United States. For ease of use, the values in this table are repeated in Section 4.0.

Table 3-4.   Quantiles of the estimated PM_{coarse} to $PM_{2.5}$ ratio parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
K	0.28	0.37	0.46	0.56	0.72	0.87	1.04	1.29	1.6	2.22	3.29

3.5   Phase Shift Between $PM_{2.5}$ and PM_{coarse} Cycles

This parameter controls the difference in time between the peak of $PM_{2.5}$ in a year and the
peak of PM_{coarse} in a year. The units of the parameter are "months." This parameter is estimated
by first calculating the mean $PM_{2.5}$ and PM_{coarse} levels for each month in the dataset for a site.
Next, the month with the highest mean level is found for each series. Subtracting the number of
the month with the highest average PM_{coarse} level from the number of the month with the highest
average $PM_{2.5}$ level yields the estimate. For instance, if the month in the dataset with the highest
average $PM_{2.5}$ level is August and the month in the dataset with the highest average PM_{coarse} level

is June, then the estimate is +2 months. Since the sine wave is cyclical, one may add or subtract 12 from the value obtained without changing the sine wave produced. In other words, if PM_{coarse} peaks in December and $PM_{2.5}$ peaks in January, the estimate is 1 - 12 = -11. This value is equivalent to a value of 1. For consistency, this document reports equivalent values between -5 and 6. Use of equivalent values between -5 and 6 is not required in the software tool. The estimated values for the shift parameter across 622 sites in the United States are presented in the following bar chart, Figure 3-1.
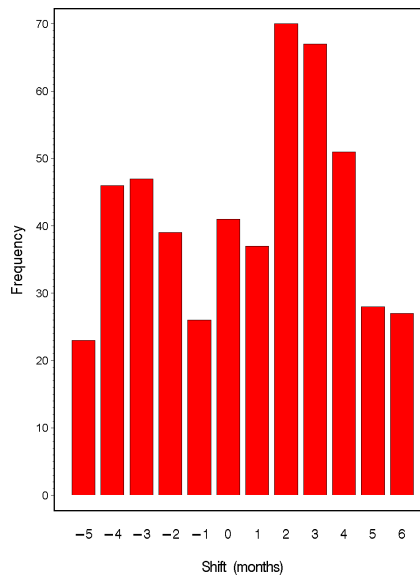

Figure 3-1.   Bar chart of the shift parameter.

3.6   Correlation Between $PM_{2.5}$ and PM_{coarse}

This parameter estimates the correlation between the two series. The estimation of this parameter can be affected by autocorrelation and seasonality, so the calculation is quite complex. First, consider only the $PM_{2.5}$ series. The series is subsetted so that it includes every 6th day measurements (deleting if needed), and the natural log of each term is taken. Next, a new sequence of numbers is created with the differences of successive pairs in the sequence of the logs. Every other term in this sequence is removed. This result is saved and the corresponding calculations for the PM_{coarse} series are performed. A single series is formed by adding the

corresponding elements of these two series. Let SS = the standard deviation of this set of numbers. In addition, let $S6_{25}$ and $S6_{coarse}$ be the values calculated in Section 3.2 for the $PM_{2.5}$ and PM_{coarse} series, respectively. The correlation is estimated by

$$\left[SS^2 - \left(S6_{coarse}^2 + S6_{25}^2\right)\right] / \left(2 \times \sqrt{S6_{coarse}^2 \times S6_{25}^2}\right).$$ Table 3-5 shows several quantiles of the estimated value of the correlation between the two series across several sites in the United States. For ease of use, the values in this table are repeated in Section 4.0.

Table 3-5.   Quantiles of the estimated PM$_{2.5}$ to PM$_{coarse}$ correlation parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
Correlation	-0.23	-0.05	0.06	0.12	0.19	0.25	0.31	0.39	0.46	0.56	0.69

3.7   Type I and Type II Error

Type I error and Type II error describe the probability of making the wrong decision under a specified set of conditions. Type I error is the probability of observing a percentile (or annual mean) above the specified daily (or annual) standard when the true ambient level (free from measurement error and bias) is below the standard. EPA QA/G-4 recommends starting with a Type I error rate of 1 percent and making adjustments as necessary. In the tool, both parameters are restricted to be at least 1 percent, as otherwise more simulations are needed to get robust results. See Step 6 of EPA QA/G-4 for additional guidance.

Type II error is the probability of making the opposite mistake: observing a value below the action limit when the true ambient level (free from measurement error and bias) is above the action limit. Since the curves show the probability of observing a value above the action limit, the value of 1 minus the Type II error is shown on the y-axis of the graphs. EPA QA/G-4 recommends starting with a Type II error rate of 1 percent and making adjustments as necessary. This parameter is also restricted to be at least 1 percent by the tool. See Step 6 of EPA QA/G-4 for additional guidance.

3.8   Annual Standard

The annual standard is the yet to be determined level of the annual NAAQS for PM_{coarse}. It is assumed that the standard will be based on the mean of three consecutive annual means.

3.9   Daily Standard

The daily standard is the yet to be determined NAAQS daily standard for PM_{coarse}. It is assumed that the standard will be based on a calculation similar to the one used for $PM_{2.5}$; namely, on the mean of the annual percentiles from three consecutive years. Further, it is assumed that the percentiles for each year will be calculated as the 98th percentile is for $PM_{2.5}$.

3.10   Percentile for the Daily Standard

The percentile for the Daily Standard is the yet to be determined percentile basis for the NAAQS or local daily standard. Under the assumptions above, the simulator will create data with the true percentile of PM_{coarse} levels equal to the percentile indicated. Then, after incorporating bias and measurement error, the DQO tool will calculate the probability of the observed value of this percentile of PM_{coarse} levels exceeding the daily standard.

3.11   1 in *m* Day Sampling

This is the sampling frequency. The value of *m* must be an integer from 1 to 12 and denotes the number of days between successive samples. 1, 3, 6, and 12 are the most common values. As an example, setting *m* to 6 produces approximately 15 sampling days each quarter. It is assumed that the $PM_{2.5}$ and PM_{10} measurements are on the same schedule. Of course, it is possible to have one monitored more frequently than the other, but only the data from the common sampling times are used for NAAQS decisions.

3.12 Completeness

Completeness is the minimum acceptable percentage of the data that is intended to be collected. Completeness is included in the DQO tool to mimic random occurrences of data loss, such as a power outage on a scheduled sampling day. The criterion is applied quarterly. Thus, if completeness is set to 0.75, the DQO tool removes 25 percent of the data from each quarter of each year. The completeness requirements are independently applied to the PM_{10} and $PM_{2.5}$ data. If a direct requirement on the PM_{coarse} completeness is desired, then the $PM_{2.5}$ completeness should be set to 1 and the PM_{10} completeness to the desired level for PM_{coarse}.

3.13 Bias

The bias input is the maximum allowable measurement bias as a proportion of the truth for $PM_{2.5}$ and PM_{10}. Bias is a consistent measurement error – a tendency to always either overestimate or underestimate the truth. For the DQO tool, bias is represented as a proportion error. If a 10 percent bias is desired in the system, enter 0.1 for the bias term. The DQO tool accepts only positive values for quantifying bias, but both positive and negative biases are simulated.

3.14 Measurement Coefficient of Variation (CV)

Measurement coefficient of variation (CV) quantifies the size of the random component of the measurement error. It is expressed as a proportion of truth for both $PM_{2.5}$ and PM_{10}. The random component of the measurement error is assumed to follow a normal distribution with a mean of 0 and a standard deviation that is proportional to the true value (for the given day). Enter 0.1 for 10 percent.

4.0 DISTRIBUTION OF ESTIMATES ACROSS THE UNITED STATES

The quantiles of the parameters estimated in the previous sections are reprinted in Table 4-1. These numbers are intended to guide users by demonstrating the range of values found across the nation. Users may consider values outside of the ranges shown here to account for the fact that the PM_{10} data available for our analysis was in standard units. In total, 622 sites across the United States were examined (refer to Figure 4-1). Sites with fewer than three observations for any single month were not considered. Also, sites where there did not exist more than one set of two consecutive days of data were removed for autocorrelation calculations since autocorrelation could not be calculated for those sites. For all parameters except the autocorrelation parameters, quantiles were calculated using 502 sites. The autocorrelation parameters were calculated using 65 sites.

Figure 4-1.   Number of sites reporting PM$_{10}$ and PM$_{2.5}$ readings by state.

Table 4-1.   Estimated particulate matter parameters

Quantile	2.5	10	20	30	40	50	60	70	80	90	97.5
$PM_{2.5}$ ratio	1.46	1.63	1.77	1.88	2.02	2.14	2.28	2.58	3.03	4.01	5.72
PM_{coarse} ratio	1.68	2.05	2.32	2.73	3.24	3.82	4.42	5.54	8.01	14.34	52.52
$PM_{2.5}$ CV	0.35	0.41	0.45	0.48	0.51	0.53	0.56	0.6	0.64	0.69	0.8
PM_{coarse} CV	0.4	0.49	0.56	0.61	0.66	0.71	0.76	0.84	0.93	1.08	1.39
$PM_{2.5}$ autocorrelation	0	0.06	0.25	0.35	0.42	0.45	0.48	0.51	0.59	0.68	0.94
PM_{coarse} autocorrelation	0	0.13	0.19	0.22	0.28	0.31	0.44	0.48	0.51	0.64	0.81
K	0.28	0.37	0.46	0.56	0.72	0.87	1.04	1.29	1.6	2.22	3.29
correlation	-0.23	-0.05	0.06	0.12	0.19	0.25	0.31	0.39	0.46	0.56	0.69

5.0 MODEL ASSUMPTIONS

In addition to several assumptions about the functional form of the PM model (e.g., sinusoidal annual patterns), the DQO tool makes two other important assumptions. First, the DQO tool assumes that the $PM_{2.5}$, PM_{coarse}, and PM_{10} readings are consistent from year to year. Second, the DQO tool assumes that the simulated values are multiplicatively scalable to higher mean values.

5.1   Year-to-Year Consistency

The assumption of year-to-year consistency means that values of the parameters are the same across all three years simulated. This assumption is placed in the model for simplicity – it is not meant to model the true behavior of PM concentrations over long periods of time. However, this assumption provides needed stability for the simulator and allows for accurate computation of percentiles.

5.2   Scalability of PM Realizations

The DQO tool is designed to evaluate the probability of an observed percentile (or the observed annual mean) exceeding an action limit for several different possible true values of the PM process. However, the simulator does not resimulate all of the data every time the true value

is altered. Instead, the DQO tool makes the assumption that increasing the mean level of one type of particulate matter increases the mean level of other types of particulate matter proportionately. Each of the parameters described in Sections 3.1 through 3.6 are estimated under this assumption.

For example, Section 3.4 describes the parameter controlling the ratio between the means of the PM_{coarse} and $PM_{2.5}$ series. If this parameter is set to 0.5, the mean level of the PM_{coarse} series is always assumed to be half the mean level of the $PM_{2.5}$ series regardless of the value the mean level of the $PM_{2.5}$ series takes for the site. This assumption may not be entirely valid – as mean levels of $PM_{2.5}$ increase, mean levels of PM_{coarse} may not increase proportionately. If it is felt that this assumption will not be met, then a "worst" case value should be used. What constitutes the worst case for any given parameter may depend on the choices for the other parameters. It is recommended that a range of different values be used wherever it is not clear what constitutes a worst case.

6.0 CONCLUSIONS AND RECOMMENDATIONS

Users of the DQO tool should keep in mind that it simulates data that follow a somewhat idealized pattern that may not correspond exactly to reality. It is intended to guide users in making policy decisions, not as a forecasting tool.

It is also recommend that users take time to become familiar with the data (possibly downloaded from AQS) before estimating parameters to be input into the DQO tool. If the data are of low quality or very incomplete, more advanced statistical modeling techniques may be desirable for estimating the values to be input into the simulator.

7.0 REFERENCES

(40 CFR 58) U.S. Code of Federal Regulations (2002). Title 40, Volume 5, Part 58." Available at http://www.access.gpo.gov/nara/cfr/waisidx_02/40cfr58_02.html.

U.S. EPA (2003). "*User's Guide for DQO Companion for PMcoarse,*" written by Battelle for the U.S. EPA, Office of Air Quality Planning and Standards, Contract No. 68-D-98-030, Work Assignment 1-01.

U.S. EPA (2002). "DQO *Companion*, Version 1.0 User's Guide," written by Battelle for U.S. EPA under Contract No. 68-D-98-030, Work Assignment 5-07.

U.S. EPA, (2000). "Guidance for the Data Quality Objectives Process (QA/G-4)." Office of Environmental Information, Report No. EPA/600/R-96/055. Available at http://www.epa.gov/quality/qs-docs/g4-final.pdf.

APPENDIX A:

STATISTICAL MODEL

APPENDIX A: STATISTICAL MODEL

This appendix gives details of the mathematical models used for the simulation of PM$_{\text{coarse}}$ observations in the DQO tool. At a single site, the statistical model used by the simulator assumes that particulate matter concentrations deviate randomly from sinusoidal curves that complete one full cycle each year. First consider the sine curves that describe the average behavior of PM$_{2.5}$ and PM$_{\text{coarse}}$ over the course of a year. These two series are described by the equations

$$\mu_{2.5}(t) = \alpha \left[1 + \beta_{2.5} \sin\left(\frac{2\pi t}{365} \right) \right] \qquad \text{(Eq. 1)}$$

and

$$\mu_c(t) = k\alpha \left[1 + \beta_c \sin\left(\frac{2\pi t}{365} + \frac{2\pi s}{12} \right) \right] \qquad \text{(Eq. 2)}$$

where t is a time variable, α controls the mean of the PM$_{2.5}$ series, k controls the mean of the PM$_{\text{coarse}}$ series, $\beta_{2.5}$ and β_c control the amplitudes of the series, and s controls the offset between the peak of the PM$_{2.5}$ and PM$_{\text{coarse}}$ series. These two equations represent sine curves that complete a single cycle between $t = 0$ and $t = 365$. In each case, the expression inside the brackets represents a sine wave that oscillates around a value of 1. The multiplicative constants outside the bracketed expressions scale the series so that the true center of oscillation is α for the PM$_{2.5}$ series and $k\alpha$ for the PM$_{\text{coarse}}$ series. In both cases, the amplitude of the sine wave is controlled by a parameter β. This parameter is a function of the ratio variable described in Section 3.1. Specifically, $\beta = (\text{ratio} - 1)/(\text{ratio} + 1)$. The expression for the mean behavior of PM$_{\text{coarse}}$ (Eq. 2) has an additional term, $2\pi s/12$, inside the sine function that allows the peak level of PM$_{\text{coarse}}$ to occur s months before the peak of PM$_{2.5}$. Negative values are permissible for s allowing the peak for PM$_{\text{coarse}}$ to occur after that of PM$_{2.5}$. A sinusoidal curve representing the non-random component of PM$_{10}$ may be calculated by adding the two expressions:
$\mu_{10}(t) = \mu_{2.5}(t) + \mu_c(t)$.

The daily deviations of particulate matter observations from these sine waves are introduced through multiplication of the daily sine wave value by a daily random component. The characteristics of this random component are designed to mimic properties exhibited by true $PM_{2.5}$, PM_{10}, and PM_{coarse} observations. For instance, the day-to-day values of $PM_{2.5}$ levels can be correlated with the day-to-day values of PM_{coarse} levels, and the simulator allows incorporation of this between-series correlation in the random component. In addition to the random multiplicative component, the simulator allows for input of measurement error and bias to account for effects resulting from inaccurate data collection.

The following sections describe the elements of the full model in more detail. First, the multiplicative random component is described. We use a log-normal distribution (Section A.1) that allows for correlation between $PM_{2.5}$ and PM_{coarse}. In addition to correlation, we incorporate autocorrelation within each series into the calculations. These two types of correlation, correlation between the series and autocorrelation within each series, are described in detail in Sections A.2 and A.3. Finally, bias and measurement error are discussed in Section A.4

A.1   Log-normal Distribution

The observations of particulate matter recorded by samplers do not exactly follow the sinusoidal curves described in Equations 1 and 2. A certain amount of random error, along with measurement error and bias, contribute to the observed readings of particulate matter concentrations. The true observed values at time t for $PM_{2.5}$ and PM_{10} can be represented by

$$PM_{2.5}(t) = \mu_{2.5}(t) \times z_{2.5}(t) \times (1 + B_{2.5}) \times (1 + x_{2.5}(t)\eta_{2.5}) \tag{Eq. 3}$$

and

$$PM_{10}(t) = \left[\mu_{2.5}(t) \times z_{2.5}(t) + \mu_c(t) \times z_c(t)\right] \times (1 + B_{10}) \times (1 + x_{10}(t)\eta_{10}) \tag{Eq. 4}$$

where $z_{2.5}$, z_c, $x_{2.5}$, and x_{10} are random components; $B_{2.5}$ and B_{10} are bias terms; and $\eta_{2.5}$ and η_{10} are measurement coefficients of variation. These two equations look confusing, but they make intuitive sense when broken down into their constituent pieces. First, consider Equation 3, the observation equation for $PM_{2.5}$ on day t.

- The first term, $\mu_{2.5}(t)$ is the sinusoidal curve from Equation 1 that represents the average behavior of PM$_{2.5}$ over the course of a year (see Eq. 1).

- The next term, $z_{2.5}(t)$ is the random component that allows the true PM$_{2.5}$ concentration to deviate from the underlying sinusoidal curve. Since this random component is multiplied by the mean value, we restrict it to have a mean of 1. This restriction means that the average behavior of the series follows the sine curve.

- The term $(1 + B_{2.5})$ introduces bias into the measurement process. A recorder that systematically adds or subtracts from the true value of PM$_{2.5}$ when reporting the observed value is biased. A recorder that does not systematically add or subtract anything from the true value of PM$_{2.5}$ is unbiased and will have $B_{2.5} = 0$. In that case, the term $1+B_{2.5}$ is equal to 1 and has no effect on the PM$_{2.5}$ reading recorded.

- The final term, $(1 + x_{2.5}(t)\eta_{2.5})$ allows measurement error to be incorporated into the system. The value $\eta_{2.5}$ is called the coefficient of variation and is calculated as the standard deviation of the random measurement errors divided by their mean. In this expression, $x_{2.5}(t)$ is a random component that takes a different value at each time step.

Equation 4 represents the observation equation for PM$_{10}$ and is almost identical to Equation 3. In Equation 4, the two added terms inside the square brackets are the sum of the means and random components for PM$_{2.5}$ and PM$_{coarse}$. The final two terms are bias and measurement error terms corresponding to the PM$_{10}$ series. The remainder of this subsection and the next two subsections discuss the choice of statistical distribution for $z(t)$ and its construction. Discussion of the bias and measurement error terms in each series can be found in Section A.4.

Equations 3 and 4 indicate that the random deviations of the series about the average sine curve will be introduced through a multiplicative factor $z(t)$. Using a multiplicative factor (instead of an additive one like in linear regression) means that when the average level of the series increases, the variability in the observations increases as well. As an example, consider one year's worth of PM$_{2.5}$ readings from a single sampler as depicted in Figure A-1. From this figure, it is apparent that at this location the variability in the observations increases in the early

winter when the $PM_{2.5}$ readings are at their greatest value. This figure also demonstrates the cyclical behavior typical of PM readings over the course of a year.
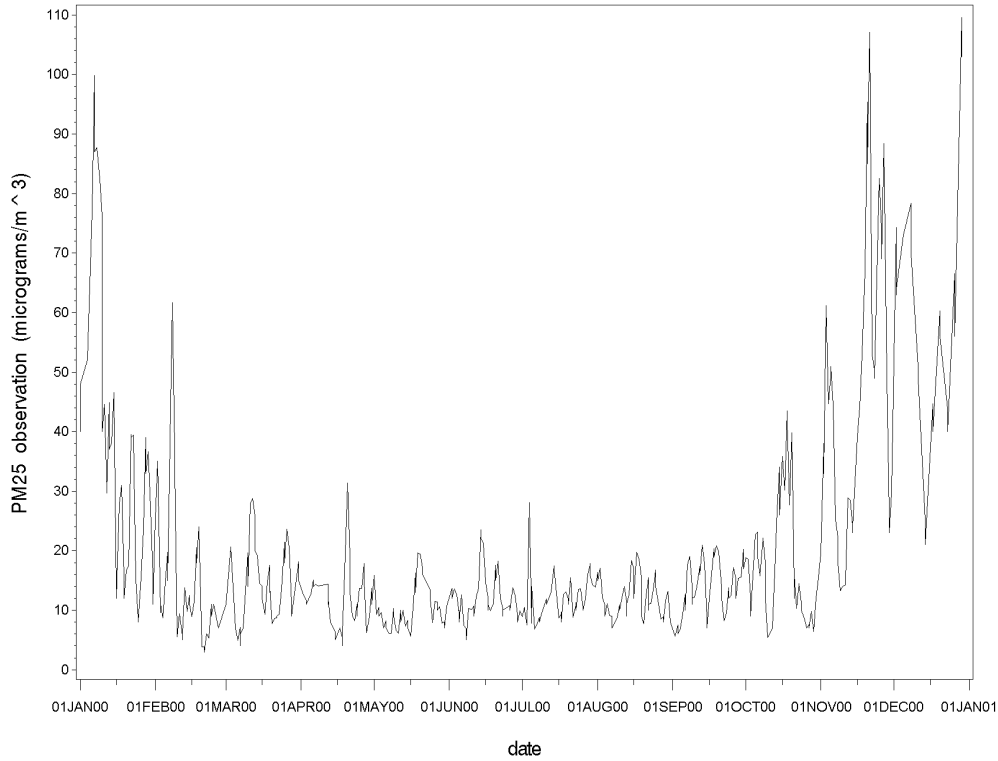

Figure A-1.   Time series of one year of fine particulate matter readings.

For the moment, assume that the random component $z(t)$ follows a log-normal distribution with a mean of 1. We will justify this assumption shortly. A log-normal random variable with a mean of 1 has the following density function:

$$f(z \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}\,z} \exp\left\{ -\frac{1}{2\sigma^2}\left(z + \frac{\sigma^2}{2} \right)^2 \right\}$$ (Eq. 5)

where σ^2 is a parameter that controls the dispersion of the distribution. Figure A-2 shows several log-normal functions for different values of σ^2. All of the densities in the figure have a mean value of 1.
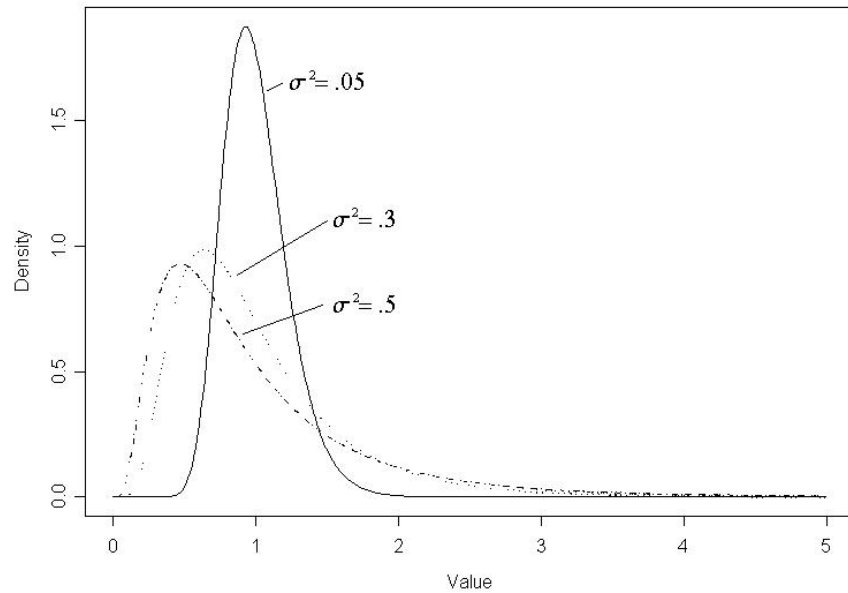
Figure A-2.   Log-normal densities with mean=1 and σ^2 = 0.05, 0.3, and 0.5.

Our model hypothesizes that, assuming bias and measurement error are nonexistent, dividing the true observation at a time step by the best fitting sine curve will produce a residual draw from a log-normal distribution. For the dataset illustrated in Figure A-1, we fit the best fitting sine curve to the data and examined the residuals. The following two figures show the original data with the fitted sine curve and the distribution of the residuals. An overlay of a log-normal distribution has been placed on top of the residuals to show that the log-normal assumption appears to be valid.
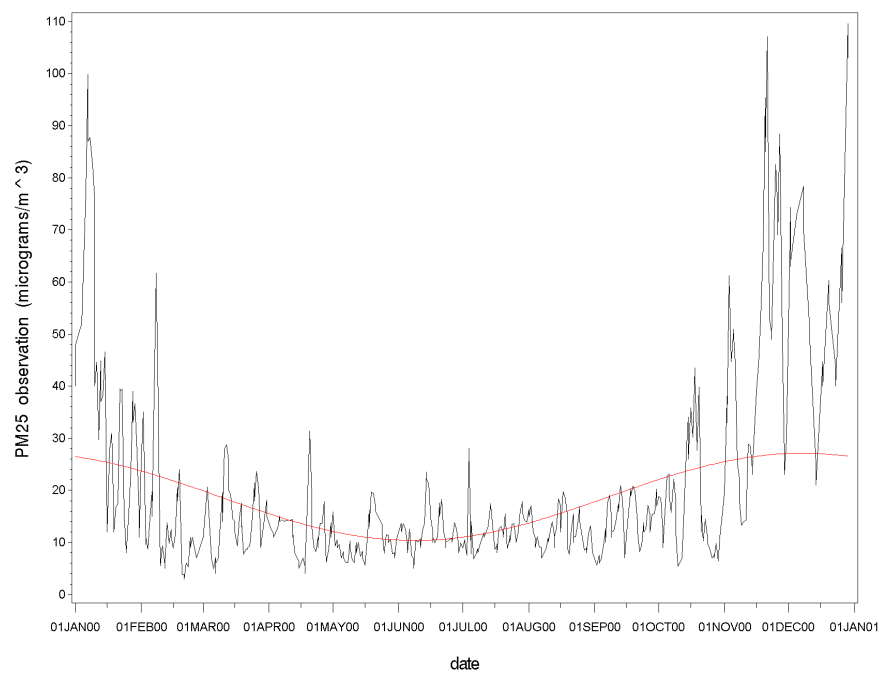
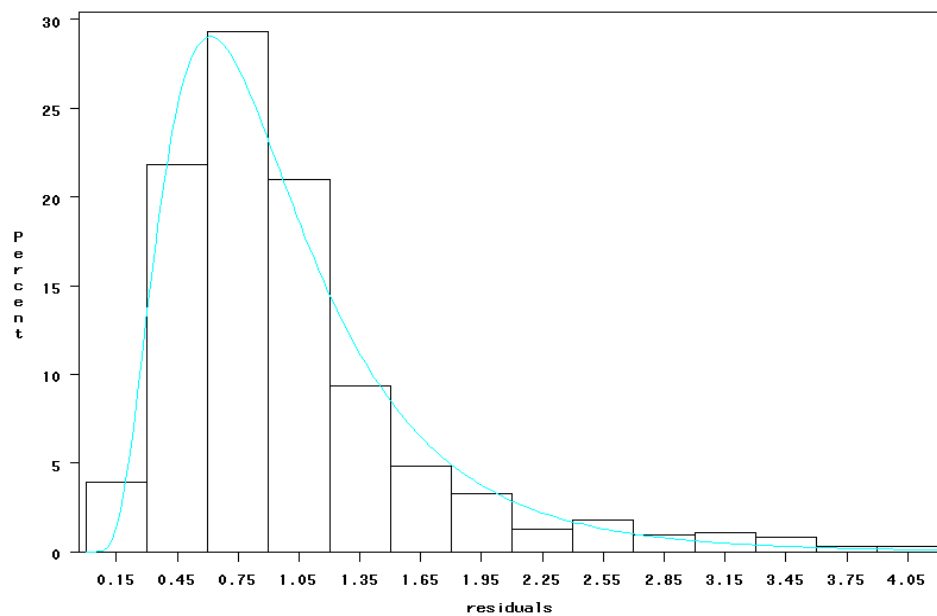**Figure A-3.   Data from Figure A-1 with the fitted sine curve (assuming log-normal multiplicative errors).**

Figure A-4.   Residuals from Figure A-3 with a log-normal density superimposed.

Similar results can be obtained for PM_{coarse} and PM_{10}. We do not demonstrate those results here.

A.2   Incorporating Correlation Between Series

The error terms for PM_{coarse} and $PM_{2.5}$ do not vary independently. When $PM_{2.5}$ rises above its mean value, PM_{coarse} is likely to do the same thing. For this reason, correlation between the two series is incorporated into the model. Bringing this correlation into the model is performed through the $z(t)$ components; when generating random log-normal draws for $z_{2.5}(t)$ and $z_c(t)$ we make sure that these draws share a specific correlation. Generation of correlated normal random variables is simple [see Robert and Cassella (1999) for details], and correlated log-normal random variables can be approximated by exponentiating correlated normal random variables.

A.3   Incorporating Autocorrelation Within Each Series

In addition to random error around the mean sine curve, it is apparent that the random components exhibit autocorrelation. In other words, for a single series, random realizations at adjacent time steps are not independent. Look again at the series in Figure A-3. When the series goes above its mean value, chances are that the value of the next random component will also be above the mean value. The series does not "jump around" enough to be considered completely random. Incorporation of autocorrelation into a single series is performed in the same manner as incorporation of correlation between the series – we introduce it through the random component $z(t)$. If the series $z(t)$ exhibits autocorrelation, the observations will as well. We generate this autocorrelation in the simulator while maintaining the overall log-normal distribution of the full set of values $z(t)$.

A.4   Effects of bias and measurement error

Bias and measurement error are introduced into the model as described in Equations 3 and 4. Bias is a consistent addition or subtraction from the true value of a process. In this

model, bias is considered to be multiplicative, so it should be interpreted as a proportion bias. In other words, if the sampler being modeled consistently underestimates the true $PM_{2.5}$ concentration by 2 percent, we could set $B_{2.5} = -0.02$. For the purposes of simulation, we assume that the true bias is unknown and that the user may be able to determine a maximum possible amount of bias, either positive or negative, for the sampler. If the user is unable to determine the maximum bias possible, we recommend using the maximum level acceptable for $PM_{2.5}$, namely 10 percent (40 CFR 58).

Measurement error is simply random error in the readings around the true value. As with bias, the model used to simulate PM assumes a multiplicative error, so measurement error should be considered a proportion error. The simulator draws the random component of the error from a standard normal distribution and multiplies it by the user-input coefficient of variation η.

REFERENCES

(40 CFR 58) U.S. Code of Federal Regulations (2002). "Title 40, Volume 5, Part 58." Available at http://www.access.gpo.gov/nara/cfr/waisidx_02/40cfr58_02.html.

Robert, C. P., and Casella, G. (1999). Monte Carlo Statistical Methods. Springer-Verlag.

Appendix 3B

Precision and Bias Estimates used in PMc Data Quality Objective Report

PMc Precision and Bias Estimates

Precision

Precision estimates for PMc were calculated for each method by site and for each method aggregated over all three sites. The following formula was implemented.

$$\text{Precision} = \sqrt{\dfrac{n\left(\sum\limits_{i=1}^{n}(CV_i)^2\right) - \sum\limits_{j=1}^{k}\left[\sum\limits_{\substack{i=1 \\ X_{ij}\ \text{not missing}}}^{n}\left(\dfrac{X_{ij}-M_i}{M_i}\right)\right]^2\left(\dfrac{1}{k-1}\right)}{n(n-1)}}\ , \quad \text{where}$$

X_{ij} = measurement from i^{th} sampling period (i = 1…n) and j^{th} sampler (j = 1…k)

$M_i = \dfrac{1}{k}\sum\limits_{j=1}^{k}X_{ij}$ is the mean of the measurements from the ith sampling period

$CV_i = S_i\big/M_i$ is the coefficient of variation from the ith sampling period

$S_i = \sqrt{\dfrac{k\sum\limits_{j=1}^{k}(X_{ij})^2 - \left(\sum\limits_{j=1}^{k}X_{ij}\right)^2}{k(k-1)}}$ is the standard deviation of the measurements from the ith

sampling period.

n = the number of times there are two or more co-located measurements available for estimating the precision. For calculating precision by site, n had a maximum of 30. For the estimates which were aggregated over sites, n represented the number of runs over the three sites (up to a maximum of 90).

Note: The estimate adjusts for bias in the individual samplers (the rightmost summations in the precision formula). However, the derivation of the formula assumes complete data and the adjustment may not be the "best" possible when some of the samplers do not operate all of the time.

Bias

Bias estimates for PMc were derived for each individual sampler by site, each sampler aggregated over all sites, each method by site, and each method aggregated over all sites. The FRM-1 sampler was chosen as the standard (truth) for the bias calculations. The following formula illustrates the calculation for each sampler.

Let X_{ij} = measurement from i[th] sampling period (i = 1…n) and j[th] sampler (j = 1…k)

Mean bias for j[th] sampler $= \left(\dfrac{1}{n}\right)\displaystyle\sum_{i=1}^{n}\left(\dfrac{X_{ij}-FRM1_i}{FRM1_i}\right)$

Mean absolute bias for j[th] sampler $= \left(\dfrac{1}{n}\right)\displaystyle\sum_{i=1}^{n}\left(\dfrac{|X_{ij}-FRM1_i|}{FRM1_i}\right)$

In computing the mean bias and mean absolute bias for each method, the mean was taken over all samplers of the same type. As with the precision estimates, n varied according to the level of aggregation.

Since the mean absolute bias statistic uses absolute values, the estimate does not have a tendency (negative or positive) associated with it. A sign will be designated by rank ordering the relative percent differences (with signs) for all bias estimates for a particular sites. Calculate the 25th and 75th percentiles. The absolute bias would be flagged as positive if both the 25th and 75th percentiles where positive and negative if both are negative. The mean absolute bias would not be flagged with a sign if the percentiles were different signs.