SIMULATION OF RAW WATER AND TREATMENT PARAMETERS IN SUPPORT OF THE DISINFECTION BY-PRODUCTS REGULATORY IMPACT ANALYSIS

June 10, 1992

Frank J. Letkiewicz
Abt Associates, Inc.

William Grubbs and Mike Lustik
Science Applications International Corporation

John Cromwell, Jeff Mosher and Xin Zhang
Wade Miller Associates, Inc.

Stig Regli
U.S. Environmental Protection Agency
Office of Ground Water and Drinking Water

1. INTRODUCTION

In a related paper, Cromwell et al. (1992) addresses the many analytical complexities and uncertainties that EPA decision-makers are facing in their efforts to delineate, measure, and balance the cost and health implications of regulatory alternatives being considered to control the levels of disinfection by-products in public drinking water supplies. To assist the agency's decision-makers in grappling with these complexities and uncertainties, the EPA's Office of Ground Water and Drinking Water has undertaken an effort to model how the water supply industry may respond to possible rules, and how those responses may affect human health risk with respect both to the toxicity of the disinfection by-product chemicals and to the incidence of waterborne disease due to microbiological contaminants. The model is referred to as DBPRAM – the Disinfection By-Product Regulatory Analysis Model.

The DBPRAM has three main components. The first component involves the creation of sets of simulated water supplies that are intended to be representative of the range of conditions (and combinations of conditions) currently encountered by public water supplies with respect to certain raw water quality and water treatment characteristics. The raw water and water treatment characteristics described are those that both influence by-product formation potential and constrain alternatives available for modifying existing water treatment practices to meet regulatory goals. This paper is concerned primarily with presenting and discussing the methods, underlying data, assumptions, limitations, and results for this first part of the DBPRAM model.

The "output" from the first component of the model (namely the sets of simulated public water supplies) constitutes the "input" to the second component of the model, wherein compliance choices are simulated for these water supplies to meet alternative DBP regulatory constraints. At the heart of this second component of the DBPRAM is the water treatment plant (WTP) model. The WTP model is designed to simultaneously calculate the concentration of disinfection by-products formed and disinfection levels achieved in a water supply with specified raw water quality and existing water treatment characteristics. The WTP model is also designed to ensure that the water supply meets disinfection constraints set by the Surface Water Treatment Rule, taste and odor constraints, and corrosion control constraints set by the lead rule. The formation equations and operational aspects of the WTP model have been described in detail by Harrington et al. (1991).

Gelderloos et al. (1992) describes the specific application of the WTP model to simulate compliance choices as part of the DBPRAM. In the WTP part of the DBPRAM, each of the simulated public water supplies is evaluated to determine whether it meets all of the established constraints, including those being considered for the by-products (i.e., meeting target MCLs). If the DBP goals are not met, the WTP model allows for a sequential evaluation of water treatment modifications to determine the least cost approach to meeting all of the desired by-product, disinfection, and other water treatment constraints.

The third main component of the DBPRAM involves an aggregate risk assessment based on the predicted levels of disinfection by-products and microbiological organisms in the distribution systems of the simulated water supplies following the treatment selected in the second part of the model. This risk assessment focuses on the estimation of annual cancer cases associated with ingestion of certain trihalomethane and haloacetic acid by-products, and on both endemic and outbreak cases of giardiasis for the various regulatory alternatives under consideration.

Again, this paper focuses on the first of the three components of the overall DBPRAM model. The remainder of this paper has two main parts. The first part provides a discussion of the general methodology and assumptions used to create the simulated water supplies. The second part provides specific information on the data used and results obtained for each of the raw water quality and water treatment characteristics simulated.

2. GENERAL METHODOLOGY AND ASSUMPTIONS

For the DBPRAM modeling effort, the universe of public water supplies has been stratified into several groups based on size, water source, and existing treatment characteristics. First, there has been a distinction made between large and small water systems on the basis of the size of the population served. Those serving 10,000 or more people are designated as large systems; those serving fewer than 10,000 are considered small. To date the DBPRAM modeling effort has focused on the large systems. These large systems have been further stratified into six major groups reflecting water source and certain existing treatment conditions. As shown in Exhibit 1, the six strata are:

Ground water without disinfection;
Ground water with disinfection and softening;
Ground water with disinfection, but without softening;
Surface water without filtration;
Surface water with filtration and softening; and
Surface water with filtration, but without softening.

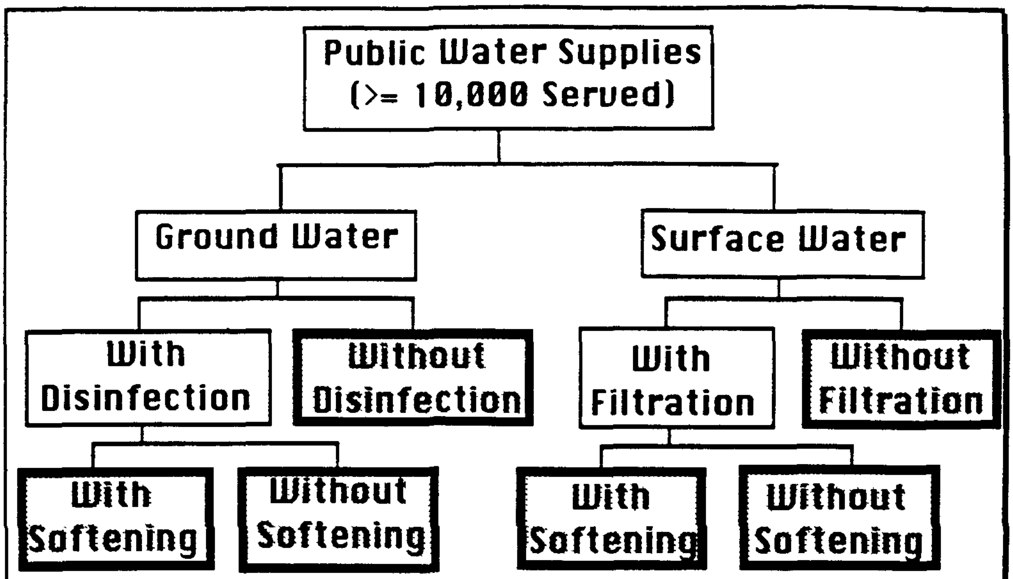


Exhibit 1. Stratification of Large PWSs for the DBPRAM

The stratification of large public water supplies into these six groups primarily reflects the different treatment process characteristics inherent to each, and the influence of those processes on by-product formation. It also reflects differences in the distributions of some of the raw water quality characteristics among these strata, differences that in part account for the different kinds of treatment used.

To date, the WTP model has been developed to an operational level only for one of these six treatment strata, namely the filtered surface water systems without softening, which are subsequently referred to here as the "SNS" systems. The DBPRAM modeling effort conducted thus far has, therefore, been limited to the consideration of this SNS group, and Section 3 of this paper, which addresses specific results of the raw water and treatment parameter simulation modeling, deals only with this SNS group. This strata represents water treated and distributed to approximately 103 million people.

For input to the WTP component of the model, 100 simulated SNS water supplies were created using Monte Carlo simulation techniques. Each of the 100 simulated supplies is defined by a unique set of values for eleven raw water quality variables and five water treatment characteristic variables assigned through the simulation procedure. These sixteen variables are identified in Exhibit 2.

As noted previously, the WTP model provides for the calculation of disinfection by-product levels at various points in the treatment and distribution systems. More specifically, the WTP model is used to estimate the concentrations of four trihalomethanes and five haloacetic acid by-products.

Exhibit 2. Raw Water Quality and Water Treatment Variables Defining the SNS Group of Simulated Water Supplies

	Units	Included in DBP Formation Equations?	Independent or Dependent Variable?
Raw Water Quality Variables			
pH	unitless	Yes	Independent
Total Organic Carbon (TOC)	mg/L	Yes	Independent
Bromide	mg/L	Yes	Independent
Average Temperature	^oC	Yes	Independent
UV-254	cm^{-1}	Yes	Dependent on TOC
Total Hardness	mg/L ($CaCO_3$ equivalents)	No	Independent
Turbidity	Turbidity Units	No	Independent
Ammonia	mg/L	No	Independent
Alkalinity	mg/L ($CaCO_3$ equivalents)	No	Dependent on pH
Calcium Hardness	mg/L ($CaCO_3$ equivalents)	No	Dependent on Total Hardness
Minimum Temperature	^oC	No	Dependent on Average Temperature
Water Treatment Variables			
Distribution Residence Time	Days	No	Independent
Average Daily Flow	MGD	No	Independent
Prechlorination	–	No	Independent
Lime Dose	mg/L	No	Independent
Alum Dose	mg/L	No	Dependent on Alkalinity

The key by-product concentration that is calculated in the WTP portion of the model is total trihalomethanes (TTHM), using the following equation developed by Amy, Chadik and Chowdhury (1987):

$$TTHM = \mathbf{0.00309}[(TOC)(UV\text{ - }254)]^{0.440}(Cl_2)^{0.409}(t)^{0.265}(T)^{1.06}(pH - \mathbf{2.6})^{0.715}(Br + \mathbf{1})^{0.036}$$

where TTHM is in μmole/L, TOC is the total organic carbon in mg/L, UV-254 is the absorbance of ultraviolet light at 254 nm wavelength (in cm^{-1}), Cl_2 is the chlorine dose in mg/L, t is the reaction time in hours, T is the temperature in degrees centigrade, and Br is the bromide concentration in mg/L. The estimation of the proportion of TTHMs that is made up by each of the four individual THMs (i.e., chloroform, bromodichloromethane, dibromochloromethane, and bromoform) is based on a set of similar equations. As indicated by Gelderloos et al. (1992), the WTP model currently uses the specific THM formation equations only to estimate the relative proportions of each present. Those proportions are then combined with the TTHM value from the above formation equation to arrive at estimates for the concentrations of the individual species.

Also as indicated by Gelderloos et al. (1992), the concentrations of the individual haloacetic acids (HAAs) are determined from statistical correlations between THMs and HAAs developed by Haas (in Patania, 1991).

As can be seen from the above formation equation, only five of the sixteen variables included in the simulated water system data sets are incorporated explicitly in the TTHM formation equation. These are: UV-254, TOC, Temperature, pH and Bromide. (The time variable, t, is related to the water treatment variable of Distribution Residence Time.) The remaining water quality and water treatment variables are incorporated into separate sets of equations based upon the specific treatment being modeled for that water supply as well as the other regulatory and treatment constraints to be met as noted previously. Those treatment-related equations result in sequential changes in the values for the primary THM formation equation variables at various points in the treatment process, characterizing the dynamics of by-product levels observed through the treatment plant and distribution system of a water supply.

As discussed in more detail in sections 3.1 to 3.16 below, certain of the water quality and water treatment parameters were treated as independent variables, while others were treated as dependent variables. For the independent variables, values were selected randomly using Monte Carlo procedures, and those values were not affected by the values for any other variable. For dependent variables, the value estimated was dependent upon the value obtained for some other variable in the data set. Exhibit 2 indicates which variables were treated as independent and which as dependent, noting the other variable upon which values for the latter were based.

The following sections provide a general discussion of the sources of information used to support creation of the data sets (2.1), the methods used to obtain values in each data set for the independent variables (2.2), and the methods used to obtain values in each data set for the dependent variables (2.3).

2.1 Information Sources Used

The primary source of data used to create the simulated set of SNS water supplies was the Water Industry Data Base (WIDB). The Water Industry Data Base was initiated through a joint effort by the American Water Works Association (AWWA) and the

American Water Works Association Research Foundation (AWWARF). The main functions of WIDB are to support the AWWA and others in assessing the impacts of regulatory and legislative efforts, to assist AWWARF in focusing its research activities, and to support educational endeavors of AWWA and other interested parties.

The data collection effort for the WIDB was conducted in 1989-1990. Questionnaires were sent to all of the approximately 600 public water systems serving over 50,000 people. The response rate was better than 80 percent, resulting in data in the WIDB for approximately 500 systems. The questionnaire was aimed at obtaining information on a wide spectrum of financial, administrative, and operational characteristics of these supplies. Included among the fields of information sought in this questionnaire were several on raw water quality and existing water treatment practices.

It was recognized at the earliest stages of the DBPRAM modeling effort that WIDB would likely be relied upon heavily as a source of information, especially for the large surface water systems. This reflected the fact that, in comparison with the few other sources of information available, the WIDB contains data for a substantial portion of these supplies from across the nation for most of the variables of interest.

Depending upon the particular variable, the WIDB typically provided data for 200 to 300 SNS systems,. For the SNS group, the WIDB served as the source of data for all of the water quality and water treatment variables with four exceptions: ammonia, bromide, UV-254, and calcium hardness. The data sources used for these four variables are described below.

The ammonia data for the SNS group was obtained from the 1991 AWWARF Disinfection Survey (DS). The 1991 Disinfection Survey was conducted by the Disinfection Committee of the AWWARF's Water Quality Division. The 1991 survey is a follow-up survey to a similar disinfection survey conducted by AWWARF in 1978. The primary purpose of the survey was to document current water industry practices regarding disinfection. In the 1991 Disinfection Survey, 283 utilities responded to questions concerning current disinfection practices and relevant water quality characteristics. (Note: Although the DS data were used only for ammonia in the simulation analysis, some data are provided in DS for all of the other water quality parameters as well, with the exceptions of bromide and UV-254.)

Bromide data for the SNS group was obtained from the James M. Montgomery (JMM) case studies. The JMM study was a cooperative effort between the Association of Metropolitan Water Agencies (AMWA) and the US EPA to study the formation and control of disinfection by-products (DBPs) in drinking water systems. The study was performed by the Metropolitan Water District of Southern California and James M. Montgomery, Consulting Engineers, Inc. The two-year study (1988-1989) focused on the identification of DBPs as a function of source water quality, water treatment process selection and operation, and disinfection processes and chemicals. Thirty-five utilities were sampled in the study.

UV-254 data for the SNS group was based on information provided by the Technical Services Division (TSD) of EPA's Office of Ground Water and Drinking Water. TSD conducted two DPB field studies between 1987 and 1989 on 53 water utilities. The information collected in these studies addressed disinfection processes, other water treatment processes, and influent and effluent water quality parameters.

The information source used for estimating the relationship between calcium hardness and total hardness was a draft version of the recently completed lime softening survey sponsored by the AWWA.

2.2 Methodology for Obtaining Data Values for Independent Variables

As noted previously, most of the variables included in the data sets for the 100 simulated SNS water supplies were treated as independent variables. That is, the values assigned to those variables were not affected by the value for any other variable. The procedures used to obtain the values for the independent variables in each data set involved two basic steps:

1. Select a probability density function (PDF) to characterize the distribution of those data;

2. Use that PDF and the Monte Carlo procedures to obtain the 100 random values for that variable.

2.2.1 Selection of a Probability Density Function

There was an *a priori* determination made that the distribution of the independent variables would be described by either a normal or lognormal probability density function (PDF). Probability density functions other than the normal and lognormal could have been evaluated for each variable within each strata, and the decision to limit consideration only to those two forms was made largely to simplify the analysis. This limitation is considered reasonable, however, given the types of data being evaluated. The normal PDF and the lognormal PDF are both frequently used to describe the distributions of environmental measurements such as those included here for the raw water quality parameters.

The normal distribution is the most frequently encountered PDF for describing the random variability observed in populations and sample data, and often serves as a default assumption. The normal distribution is recognizable as the symmetrical "bell-shaped curve," in which the central value is described by the mean of the population and the dispersion around that mean is described by the standard deviation.

The lognormal distribution is frequently used when measurement data suggest a "right skewness" due to the observation of a number of high values in the data set. As noted by Travis and Land (1990) the assumption that environmental data are lognormally distributed is fairly universal Helsel (1990), discussing various distributions in the context of trace substances in the environment, noted that the lognormal distribution can mimic the shape of right-skewed data over much of the distribution even though the data are not truly lognormally distributed. The lognormal distribution is closely related to the normal distribution in that when the values for a lognormally distributed population are transformed to their logarithms, those logarithmic values are normally distributed.

The selection of the normal or lognormal distribution to characterize the data for a particular variable was made primarily through the use of a standard "goodness of fit" test performed by SAS statistical software. The test statistic used, called the Shapiro-Wilk statistic, provides a measure of the fit of a data set with the assumption that those data are a sample taken from a normally distributed population. Both the original values and the log-transformed values of the data available for each variable were tested, and generally the form showing the best fit was selected to characterize that variable. It is important to note that the "best" fit between the two distribution forms did not necessarily indicate a "good" fit.

In some cases, other factors were also considered in selecting the distributional form to use, and those are discussed later for the specific variables involved.

In addition to selecting the distributional form to be used for a particular variable, it was also necessary to estimate the parameters defining that distribution. The specific shape of either the normal or lognormal distribution is determined by two parameters usually designated as μ and σ. The first of these parameters, μ, describes the central tendency of the values for that population. For a normally distributed population, μ is the population mean. Similarly, for a lognormally distributed population, μ is the population mean of the log-transformed values, referred to as the log mean.

The second parameter, σ, describes the dispersion of the values for the population about the central tendency. For a normally distributed population, σ is the population standard deviation; for a lognormally distributed population, σ is the population standard deviation of the log-transformed values, referred to as the log standard deviation.

There are two approaches frequently used to estimate these distributional parameters: the maximum likelihood method and the regression method. It can be shown that the maximum likelihood method for estimating the parameters of a normal distribution from a sample of data reduces to the relatively straightforward computation of the sample mean and sample standard deviation.[1] Largely because of time constraints, this was the method used to estimate the distributional parameters for this analysis.

The second approach, the regression method, has been found to have advantages over the maximum likelihood method to estimate distributional parameters for environmental data (see, for example, Helsel and Gilliom, 1986). This issue is discussed further in Section 4 and, as noted there, subsequent iterations of this modeling effort will probably use the regression approach to estimate the distributional parameters.

2.2.2 Selection of Values for the Independent Variables

The Monte Carlo method for selecting values for an independent variable for each of the 100 simulated water supplies is very similar to what would occur if one were to conduct a random sampling of a population of actual supplies to obtain representative data for those variables. The population in this case exists, however, in the form of the PDF describing the distribution of values that variable may take on.

Functionally, the Monte Carlo procedure makes use of the "area under the curve" of the PDF. By definition, the total area under the curve of any PDF has a value of 1. The probability that a variable will have a value that falls within some specified range is determined from the area under the curve between the bounds of that range.

A randomly selected number between 0 and 1 can be used to determine a value for the variable being considered. Specifically, the random number is used to represent the area under the curve from negative infinity to that point. Using that value and the inverse of the standard normal distribution function, the corresponding "z-score" is determined. Then, from that z-score, and the estimated parameters of the distribution, a value for the variable is obtained using the relationship[2]:

[1] This is true when there are no censored (non-detect) values in the data, which was the case here.

[2] The symbols $\hat{\mu}$ and $\hat{\sigma}$ refer to estimates of the distributional parameters to differentiate them from the "true" values for those parameters, μ and σ, which are unknown.

$$z_x = \frac{x - \hat{\mu}}{\hat{\sigma}}$$

For example, if a variable is determined to be normally distributed with parameters for that distribution of $\hat{\mu} = 10$ and $\hat{\sigma} = 3$, a value could be obtained by choosing a random number between 0 and 1 of, say, 0.7734. Using the inverse of the standard normal distribution function, it is determined that the z-score associated with 0.7734 is 0.75. Then, using the relationship shown above and solving for x, the randomly selected value for this variable would be 12.25.

In the case of the variables using the lognormal PDF, each "x" value is calculated in terms of its logarithm, since $\hat{\mu}$ and $\hat{\sigma}$ are in logarithmic form, which is then retransformed by exponentiating to obtain the actual value input to the data set.

The actual process used to perform the Monte Carlo simulation procedure (i.e., selecting the random numbers and deriving the corresponding values for each variable) was performed using SAS statistical software.

Using the Monte Carlo procedure to obtain a sufficiently large number of random values for each independent variable, a simulated data set is created that is, generally, representative of the underlying population distribution from which that data set has been derived. Again, this is comparable to how physical sampling and analysis to obtain values from actual water supplies would also provide a representative data set.

There is one important difference that exists between the statistical sampling of a probability distribution via the Monte Carlo technique and the actual sampling of water supplies for values. By their nature, the normal and lognormal distributions allow for the probability, albeit usually very small, of extreme values that may not make physical sense and would not be observed in the real world. For example, it may happen that the Monte Carlo procedure results in the determination of a value for a water constituent that exceeds its solubility, or for a value that is less than zero. To correct for these eventualities, it was necessary to impose upper and lower bounds on the values that could be produced by the simulation process. If a value was selected that exceeded those bounds, that value was reassigned the value of the bound that it fell beyond.

2.3 Methodology for Obtaining Values for Dependent Variables

For several of the water quality and treatment parameters, it was recognized that the values they take on in raw water or in actual water supplies, are not entirely random, but are instead dependent upon or influenced by the value for another parameter. Where this was expected to be the case, linear regression analyses were conducted to test the strength of those correlations, and to provide a means to account for those relationships in the overall simulation analysis. Generally, the following four linear relationships were examined between an assumed dependent variable "Y" and some other variable "X" it was assumed to be dependent upon:

$$Y = a + bX$$
$$\ln(Y) = a + bX$$
$$Y = a + b\ln(X)$$
$$\ln(Y) = a + b\ln(X)$$

where a and b are, respectively, the intercept and slope obtained from the linear regression analysis. The slope and intercept are determined by the method of least squares, which provides the line that results in the minimum value for the sum of the squares of the individual deviations of the actual data points from that line.

The strength of the linear relationships developed were assessed primarily from the r^2 (coefficient of determination) value for each regression, The r^2 value for a regression, which lies between 0 and 1, provides an indication of the fraction of the overall variability observed in the dependent value that can be explained by the relationship with the independent variable. Thus, an r^2 value of 1 indicates a "perfect" correlation which explains all of the variability, while an r^2 value of 0.75 indicates that 75% of the variability observed in the dependent variable can be explained by its correlation with the independent variable, with the remaining 25% being unexplained variability, usual referred to as the "residual "

In many applications using such linear regressions, it is common to simply use the linear equation obtained to predict the value for the dependent variable associated with a particular value for the independent variable. In the method used for this simulation analysis, a different approach was used to reflect both the explained variability and the residual, unexplained variability.

The measure of the "scatter" of the original data about the least squares regression line obtained is referred to as the root mean square error or standard error of the estimate of y on x ($s_{y.x}$). This value is determined as:

$$s_{y,x} = \sqrt{\frac{(y_i - y')^2}{n}}$$

where y_i refers to the observed values for the dependent variable corresponding to the value for the independent variable, x_i; y' is the predicted value for that variable using x_i and the linear regression; and n is the number of data points used to perform the regression analysis.

The root mean square error, $s_{y.x}$, has properties similar to a standard deviation in that if one were to construct pairs of parallel lines to the regression line of y on x at respective vertical distances of $s_{y.x}$, $2s_{y.x}$ or $3s_{y.x}$ from it, one would find (with sufficiently large n values) approximately 68%, 95%, and 99.7% of the data points falling within those lines, respectively.

The correlation procedures used here were as follows. Correlations were not examined for all possible combinations of data, but only for those for which there was some reason to expect a correlation to occur. The combinations considered were:

Alkalinity as a function of pH
UV-254 as a function of TOC
Minimum Temperature as a function of Average Temperature
Calcium Hardness as a function of Total Hardness
Alum Dose as a function of TOC
Lime Dose as a function of Alkalinity

First, the four linear regression forms were developed. The "best" correlation was then selected using two main criteria. The first criterion used was the r^2 value with, of course, the preference given to the form having the value closest to 1. A second criterion used was based on a test of normality of the residual variability using the aforementioned Shapiro-Wilk statistic. The need to place some weight on having the residual fit a normal distribution relates to the method used to incorporate that residual variability into the selection of the values for the dependent variables, which is discussed below.

Once the particular linear equation form was selected, it was used to obtain a value for the dependent variable from the value for the independent variable obtained for that particular hypothetical water supply using a simulation process similar to that described previously. First, the independent variable in these relationships were obtained using the Monte Carlo procedure described above. That value for X was then used with the regression equation to obtain an intermediate value for the dependent variable, Y. That intermediate value for Y was treated as the μ parameter of a normal distribution for which the σ parameter was estimated by the value of $s_{y.x}$, obtained as explained above. The final value selected for the dependent variable for that simulated water supply was then obtained using that normal distribution, and the same Monte Carlo selection procedures described previously for the independent variables. As in the case of the independent variables, it was necessary to set lower and upper bounds on the values selected by this process to reflect "real world" limits.

3. RESULTS FOR INDIVIDUAL VARIABLES

This section presents a brief summary of the specific data sources used, assumptions employed, and other considerations involved in the selection of the values for each of the water quality and treatment variables used to create the 100 simulated SNS water supplies. Comparisons are also provided between the simulated data sets and the underlying data on which they are based.

3.1 pH

The underlying data for obtaining the pH values was the WIDB. As indicated below, the WIDB provided 302 data points for SNS systems, showing a range of 6.0 to 8.7, with comparable median and mean values of 7.7 and 7.8, respectively. pH values were selected for the 100 simulated data sets using the Monte Carlo simulation procedure based on an assumed normal distribution of values. Unlike most other variables for which a goodness of fit test was done for both normal and lognormal distributions, pH was assumed *a priori* to follow a normal distribution since it is already a log-transformed value. (That is, pH is defined as $-\log_{10}(1/[H^+])$, where $[H^+]$ is the hydrogen ion concentration in moles per liter.) The parameter estimates used for the distribution were the arithmetic mean and standard deviation from the WIDB. Lower and upper bounds of 4.5 and 10, respectively, were placed on the pH values selected for the 100 simulated supplies.

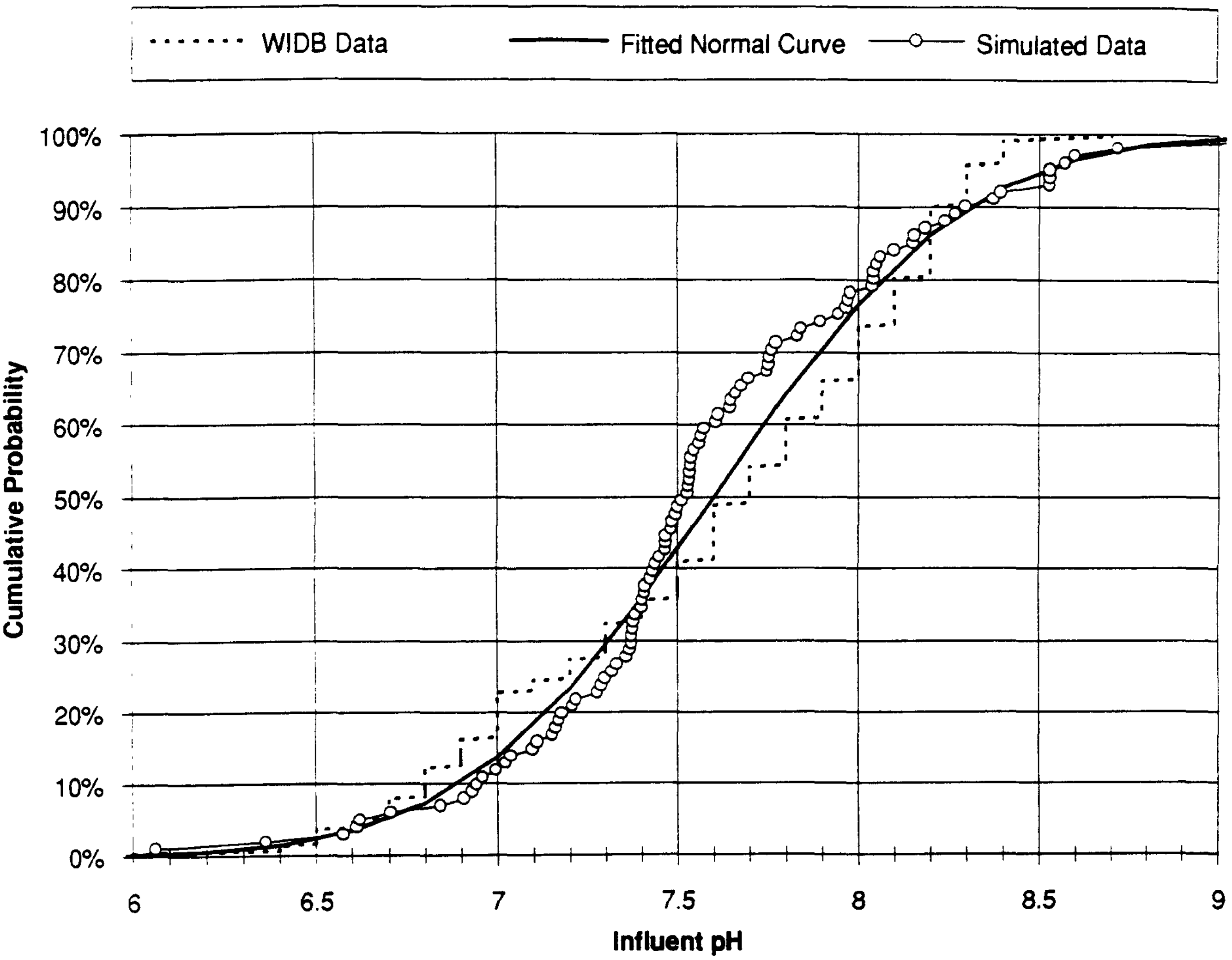
Summary Data for pH in SNS Systems

	WIDB	Simulated Data Set
N Count	302	100
Minimum Value	6.0	6.1
Maximum Value	8.7	9.1
Median	7.7	7.5
Arithmetic Mean	7.8	7.6
Standard Deviation	0.55	0.53

Exhibit 3 provides a comparison of the cumulative distribution of the WIDB data, the fitted cumulative distribution curve based on the parameter estimates made from the WIDB data, and the cumulative distribution of the 100 simulated data points.

Exhibit 3

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Influent pH



3.2 Total Organic Carbon (TOC)

As shown below, the WIDB provided TOC data for 84 SNS systems, which were used as the basis for generating the 100 TOC values for the simulated data set.

Summary Data for TOC in SNS Systems
(Concentrations in mg/L)

	WIDB	Simulated Data Set
N Count	84	100
Minimum Value	0.01 mg/L	0.39 mg/L
Maximum Value	25 mg/L	26 mg/L
Median	3.59 mg/L	3.92 mg/L
Arithmetic Mean	4.46 mg/L	5.29 mg/L
Standard Deviation	3.27 mg/L	4.92 mg/L
Log Mean	1.27	1.32
Log Standard Deviation	0.85	0.86
exp(Log Mean)	3.56 mg/L	3.73 mg/L

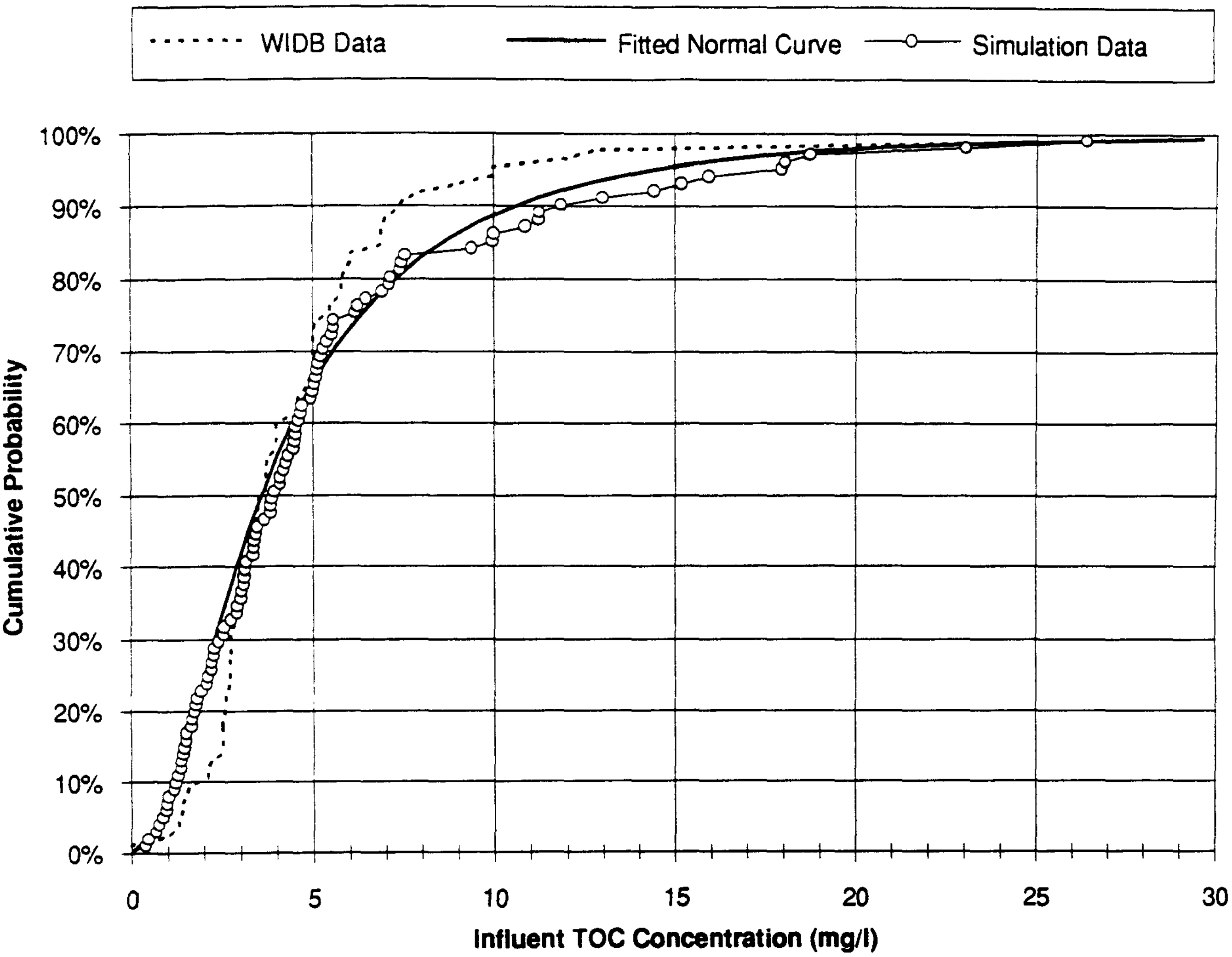
Based on the goodness of fit test performed, the lognormal distribution was selected for the TOC data. The parameter estimates for the lognormal distribution were the Log Mean and Log Standard Deviation from the WIDB data shown above. Lower and upper bounds of 0.01 mg/L and 30 mg/L were established for the TOC values selected through the Monte Carlo procedure for the 100 simulated data sets.

Exhibit 4 provides a comparison of the cumulative distribution of the WIDB data, the fitted cumulative lognormal distribution curve based on the parameter estimates made from the WIDB data, and the cumulative distribution of the 100 simulated data points. It is evident from Exhibit 4 that the derived lognormal distribution and the 100 simulated data points do not appear to compare well with the underlying WIDB data. The relative positions of these curves indicates that the derived distribution has a similar central value, but that the variance about that central value is greater than is observed in the underlying data. This results in obtaining more values in the simulated TOC data set that are further from the center of the distribution than observed in the WIDB data. For example, the derived distribution indicates that about 30% of the TOC values would fall below approximately 2.5 mg/L, while the WIDB data suggest that only about 15% fall below that value. Similarly at the high end, the derived distribution indicates that about 30% of the TOC values will exceed approximately 6 mg/L, while the WIDB data suggest that only about 20% exceed that value.

The difference between the WIDB data and the derived distributions for TOC appears to be due the presence of two "extreme" values in the WIDB data (one each at the low and high ends) that result in a higher standard deviation for the data than would be computed without those data points. As discussed further in Section 4, the influence of these data points on the shape of the derived TOC distribution would have been lessened had the regression approach been used to estimate the distributional parameters.

Exhibit 4

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Influent TOC Concentration



3.3 UV-254

Regression analyses were conducted to assess the relationship between raw water TOC and UV-254, reasoning that as TOC increased, absorbance of UV-254 would also increase. There was only a limited set of data, however, that provided both TOC and UV-254 data for the same samples, namely 23 surface water data points from the JMM study. As indicated in the preceding discussion of the methodology used to evaluate dependent variables, four linear relationships were tested and compared on the basis of r^2 values and the Shapiro-Wilk statistic (W) testing the normality of the residual variability. Shown below are the results of those tests:

$$UV\text{-}254 = a + bTOC \quad (r^2 = \mathbf{0.954}; W = \mathbf{0.969})$$

$$\ln(UV\text{-}254) = a + bTOC \quad (r^2 = \mathbf{0.625}; W = \mathbf{0.920})$$

$$UV\text{-}254 = a + b\ln(TOC) \quad (r^2 = \mathbf{0.758}; W = \mathbf{0.943})$$

$$\ln(UV\text{-}254) = a + b\ln(TOC) \quad (r^2 = \mathbf{0.791}; W = \mathbf{0.910})$$

Based on both the r^2 and the W statistic, the first of these four was used to characterize the relationship between TOC and UV-254. The parameters obtained for this relationship were:

Intercept (a):	0.03386
Slope (b):	-0.03039
Root MSE:	0.03107

Exhibit 5 displays the original JMM data and the derived linear relationship. Included on Exhibit 5 are lines showing ± root mean square error distances from the regression line. Exhibit 6 shows the scatter of the 100 simulated SNS data points selected against the obtained regression line. Lower and upper bounds of 0.01 and 1 cm^{-1} were established for the UV-254 values selected for the simulated data set.

Exhibit 5

REGRESSION ANALYSIS
Actual Data and Regression-predicted UV-254

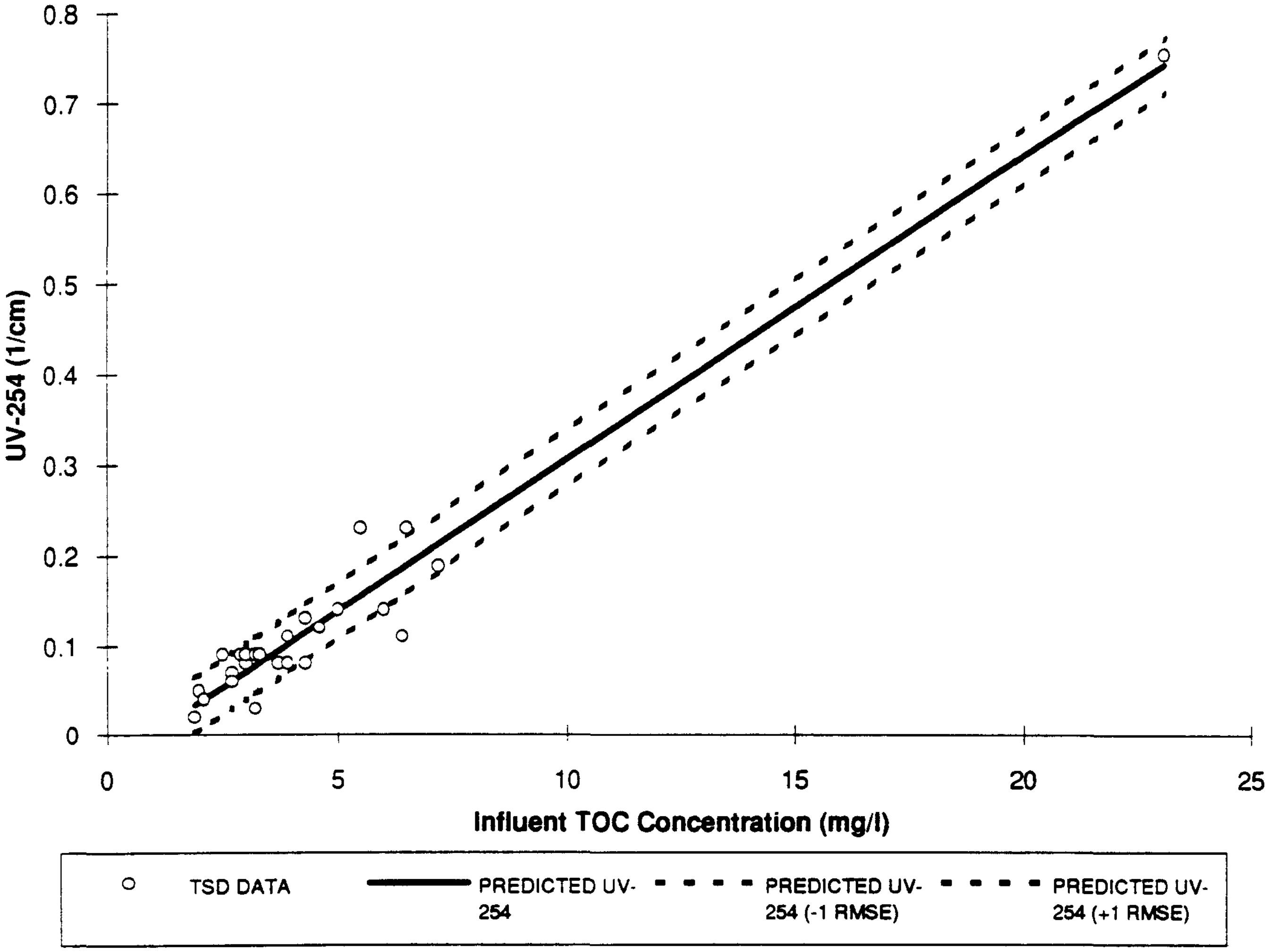
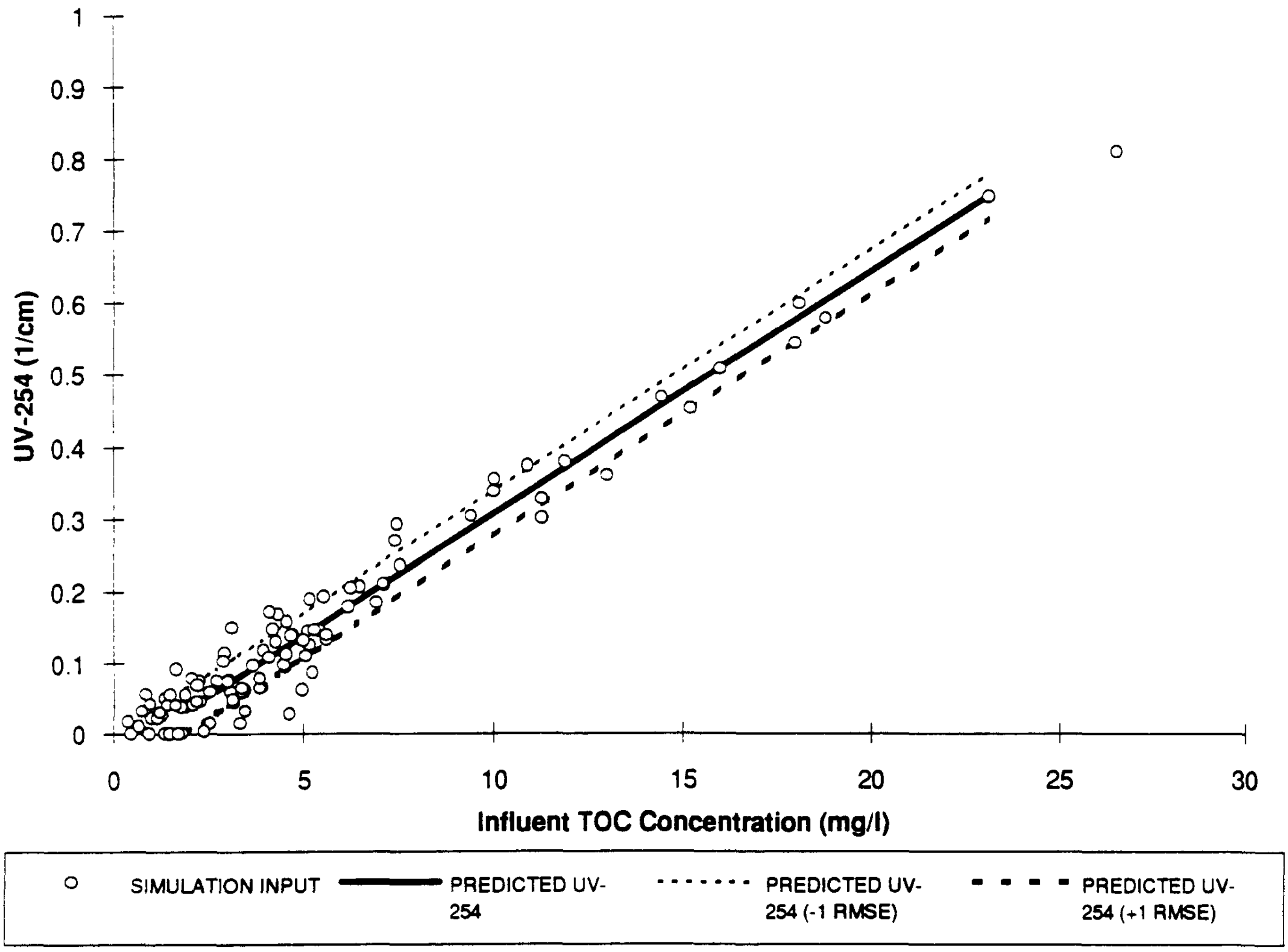


Exhibit 6

REGRESSION ANALYSIS
Simulated Data and Regression-predicted UV-254



3.4 Bromide

The available data on the occurrence of bromide in raw water used for drinking water is very limited. For the SNS systems, there were 18 data points available from the JMM study, which were used as the basis for deriving the 100 simulated data points for this analysis. Summary statistics for the JMM data and the 100 simulated data points are presented below.

Summary Data for Bromide in SNS Systems
(Concentrations in mg/L)

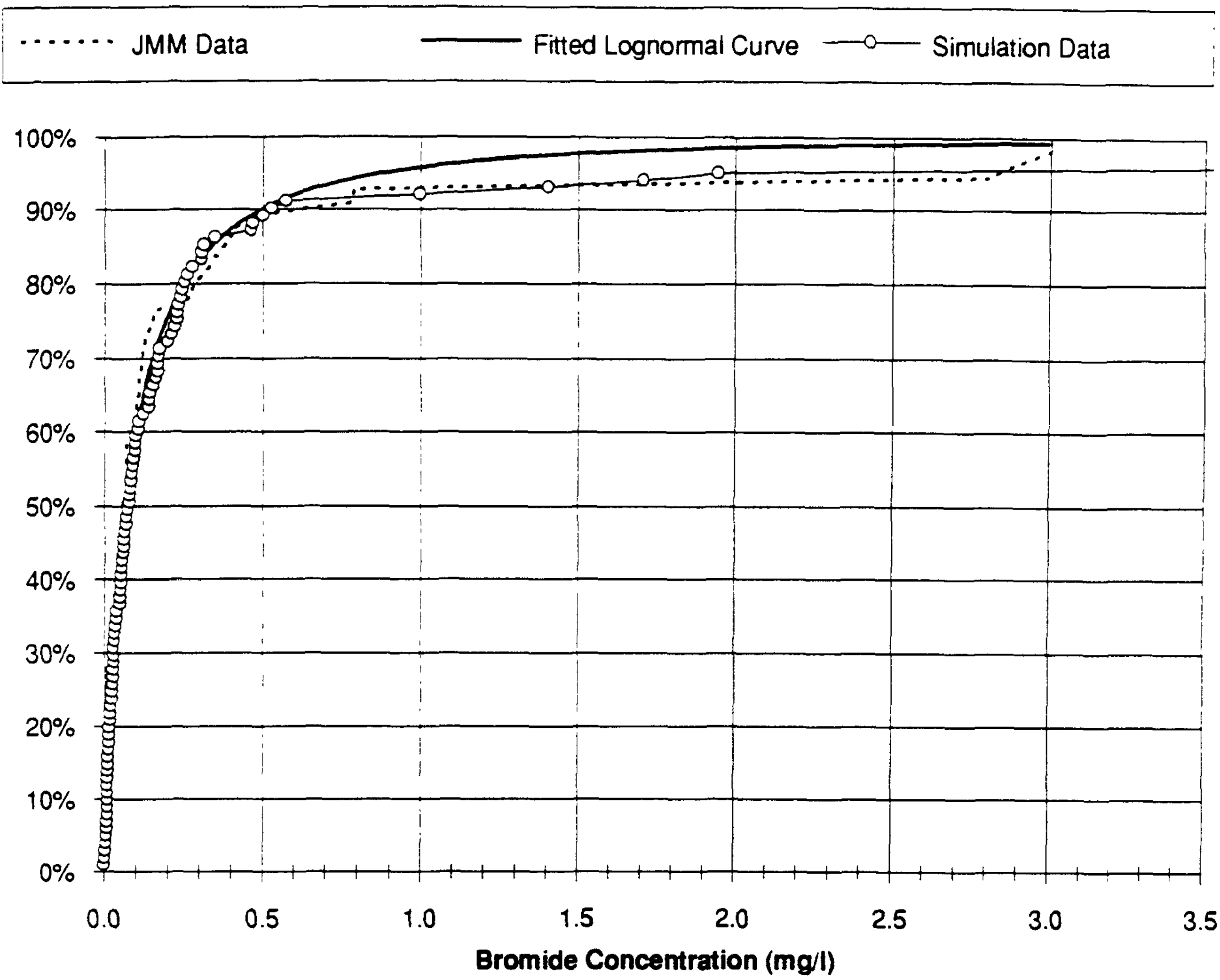
	JMM	Simulated Data Set
N Count	18	100
Minimum Value	0.01 mg/L	0.0 mg/L
Maximum Value	3 mg/L	5.0 mg/L
Median	0.06 mg/L	0.08 mg/L
Arithmetic Mean	0.27 mg/L	0.33 mg/L
Standard Deviation	0.66 mg/L	0.83 mg/L
Log Mean	-2.70	-2.58
Log Standard Deviation	1.56	1.69
exp(Log Mean)	0.067 mg/L	0.08 mg/L

Based on the goodness of fit test performed, the lognormal distribution was selected for bromide. The parameter estimates for this distribution were the Log Mean and Log standard Deviation from the JMM data shown above. Lower and upper bounds of 0 and 5 mg/L were placed on the bromide values selected for the 100 simulated data points.

Exhibit 7 provides a comparison of the cumulative distribution of the JMM data, the fitted cumulative lognormal distribution curve based on the parameter estimates obtained from the JMM data, and the cumulative distribution of the 100 simulated bromide data values. The fitted distribution and the JMM data compare well through most of the values, showing a slight divergence only at the upper tail of the curve, which may reflect the relatively small number of data points in the JMM data set.

Exhibit 7

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Bromide Concentration



3.5 Average Temperature

The average temperature data (in °C) were based on the 285 SNS data points provided in the WIDB, as summarized below.

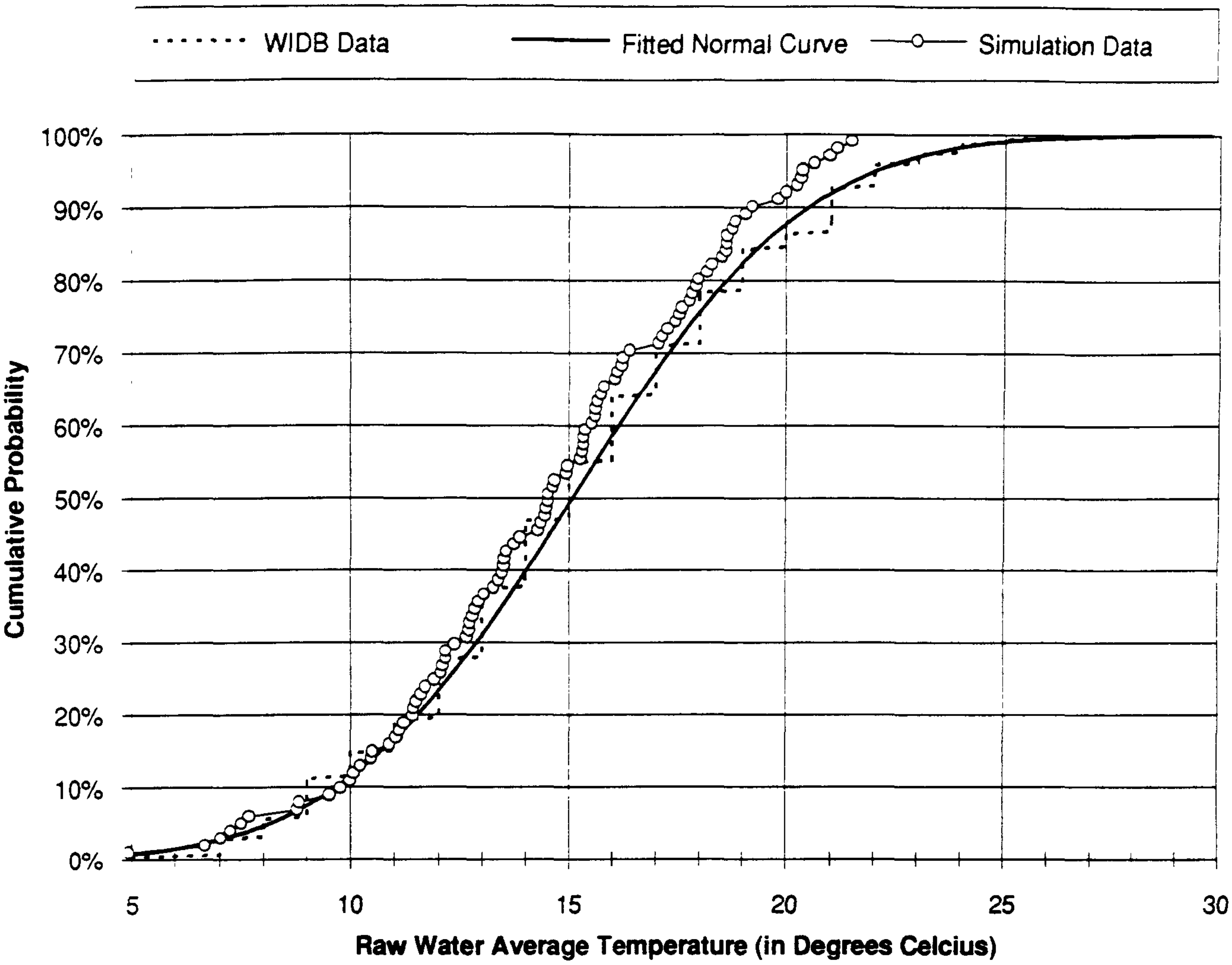
Summary Data for Average Temperature in SNS Systems
(Concentrations in mg/L)

	WIDB	Simulated Data Set
N Count	285	100
Minimum Value	5 °C	4.9 °C
Maximum Value	26 °C	21 °C
Median	15 °C	14.5 °C
Arithmetic Mean	15.1 °C	14.5 °C
Standard Deviation	4.2 °C	3.7 °C
Log Mean	2.67	3.68
Log Standard Deviation	0.299	2.64
exp(Log Mean)	14.4 °C	14.0 °C

Based on the goodness of fit test performed, the normal distribution was selected to characterize the average temperature data. Exhibit 8 compares the cumulative frequency of the WIDB data with the fitted normal distribution and the cumulative distribution of the 100 simulated data points. Lower and upper bounds of 0.5 and 30 °C were established for the 100 simulated data points.

Exhibit 8

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Raw Water Average Temperature



3.6 Minimum Temperature

It was assumed *a priori* that the Minimum Temperature of the influent raw water is related to the Average Temperature. The four regression forms were tested, the resulting statistics for which are shown below:

$$Min.Temp. = a + b\, Avg.Temp. \quad (r^2 = 0.453; W = 0.967)$$

$$\ln(Min.Temp.) = a + b\, Avg.Temp. \quad (r^2 = 0.426; W = 0.972)$$

$$Min.Temp. = a + b \ln(Avg.Temp.) \quad (r^2 = 0.398; W = 0.962)$$

$$\ln(Min.Temp.) = a + b \ln(Avg.Temp.) \quad (r^2 = 0.398; W = 0.964)$$

Although the r^2 value was highest for the Min.Temp. vs. Avg.Temp., the Shapiro-Wilk statistic for normality of the residual was higher for the **ln**(Min.Temp.) vs. Avg.Temp.; consequently this was the form chosen to describe the relationship between these variables for the SNS systems. The parameters obtained for this relationship were:

Intercept (a):	0.9334
Slope (b):	0.1478
Root MSE:	0.6167

Exhibit 9 shows the WIDB data for Minimum and Average Temperature relative to the derived relationship. Exhibit 10 similarly compares the 100 simulated data points to the derived relationship. Lower and upper bounds of 0.1 and 30 °C were placed on the simulated Minimum Temperature data points. Also, a constraint was used requiring that the Minimum Temperature value selected be less than the corresponding Average Temperature value.

Exhibit 9

REGRESSION ANALYSIS
Actual Data and Regression-predicted Raw Water Minimum Temperature

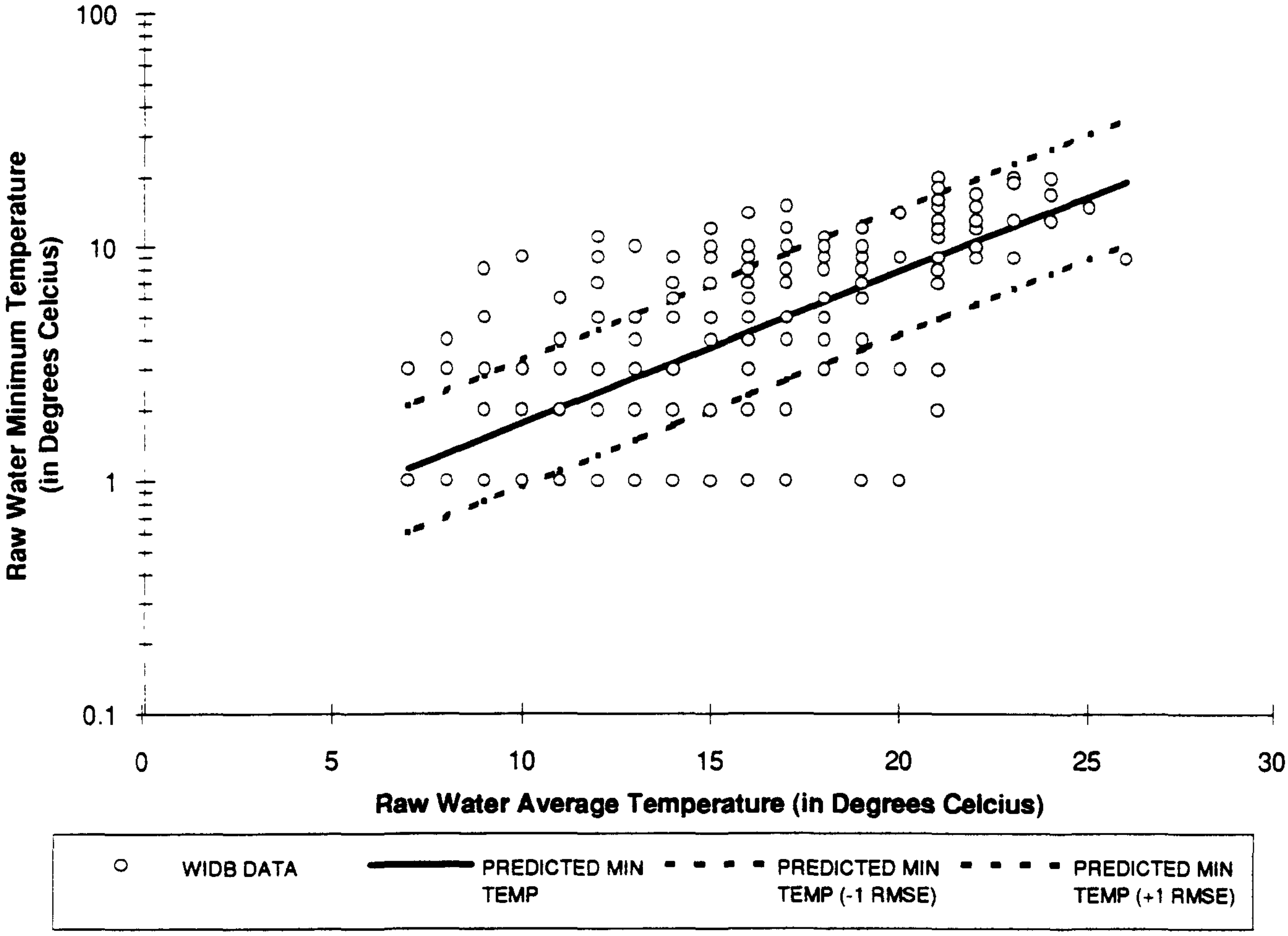
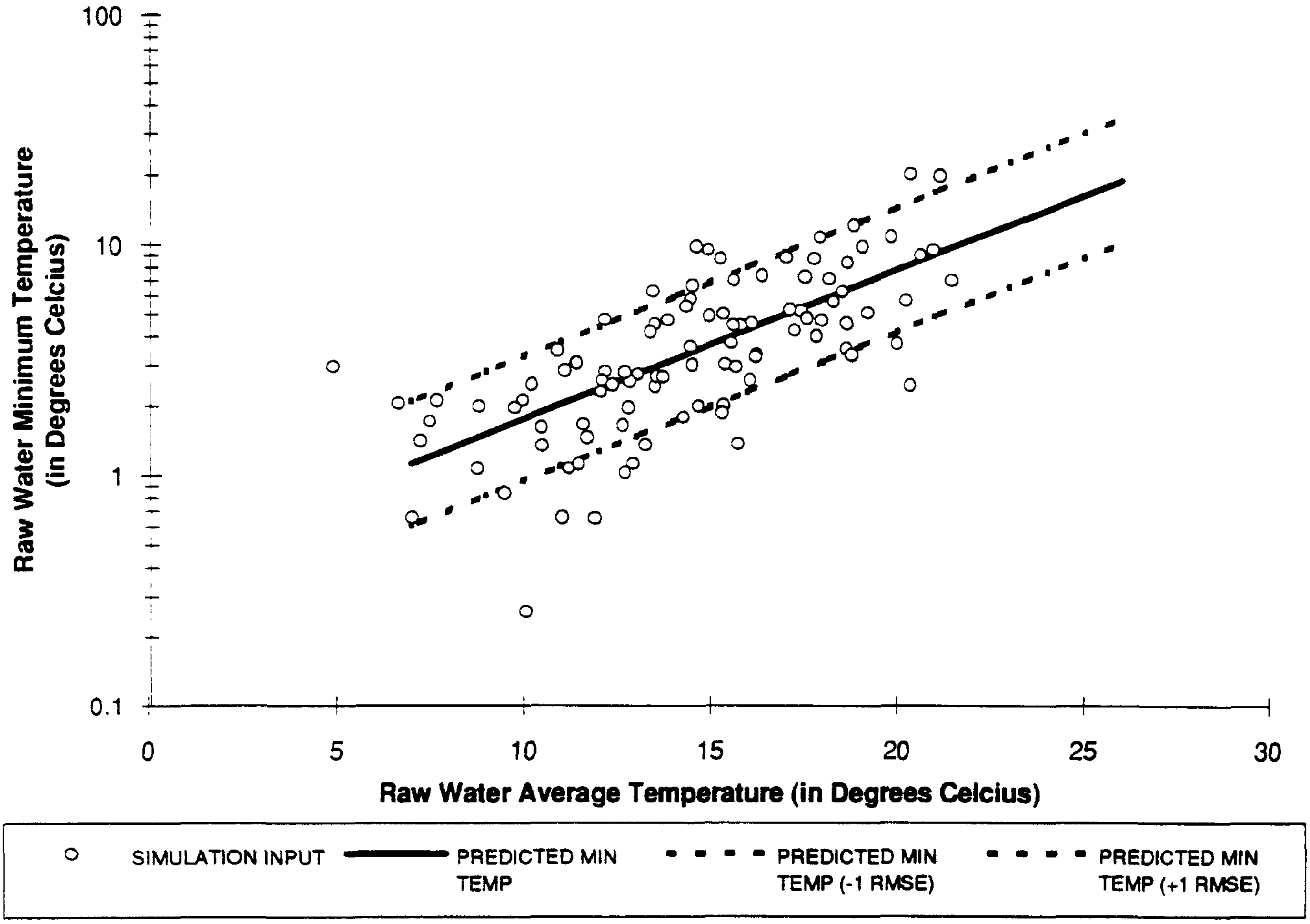


Exhibit 10

REGRESSION ANALYSIS
Simulated and Regression-predicted Raw Water Minimum Temperature



3.7 Total Hardness

As shown below, the WIDB provided 291 data points for total hardness in SNS supplies. These were used as the basis for selecting the 100 data points for the simulated SNS supplies.

Summary Data for Total Hardness in SNS Systems
(Concentrations in mg/L as $CaCO_3$ Equivalents)

	WIDB	Simulated Data Set
N Count	291	100
Minimum Value	7 mg/L	11.2 mg/L
Maximum Value	551 mg/L	779 mg/L
Median	106 mg/L	71 mg/L
Arithmetic Mean	115 mg/L	122 mg/L
Standard Deviation	86 mg/L	158 mg/L
Log Mean	4.39	4.29
Log Standard Deviation	0.92	0.98
exp(Log Mean)	80.6 mg/L	72.7 mg/L

Based on the goodness of fit test, the lognormal distribution was selected to describe the distribution of total hardness data for the simulated systems. The parameter estimates used for this distribution were the Log Mean and Log Standard Deviation from the WIDB data set shown above. Lower and bounds of 5 and 1,000 mg/L were placed on the total hardness values selected for the 100 simulated data points.

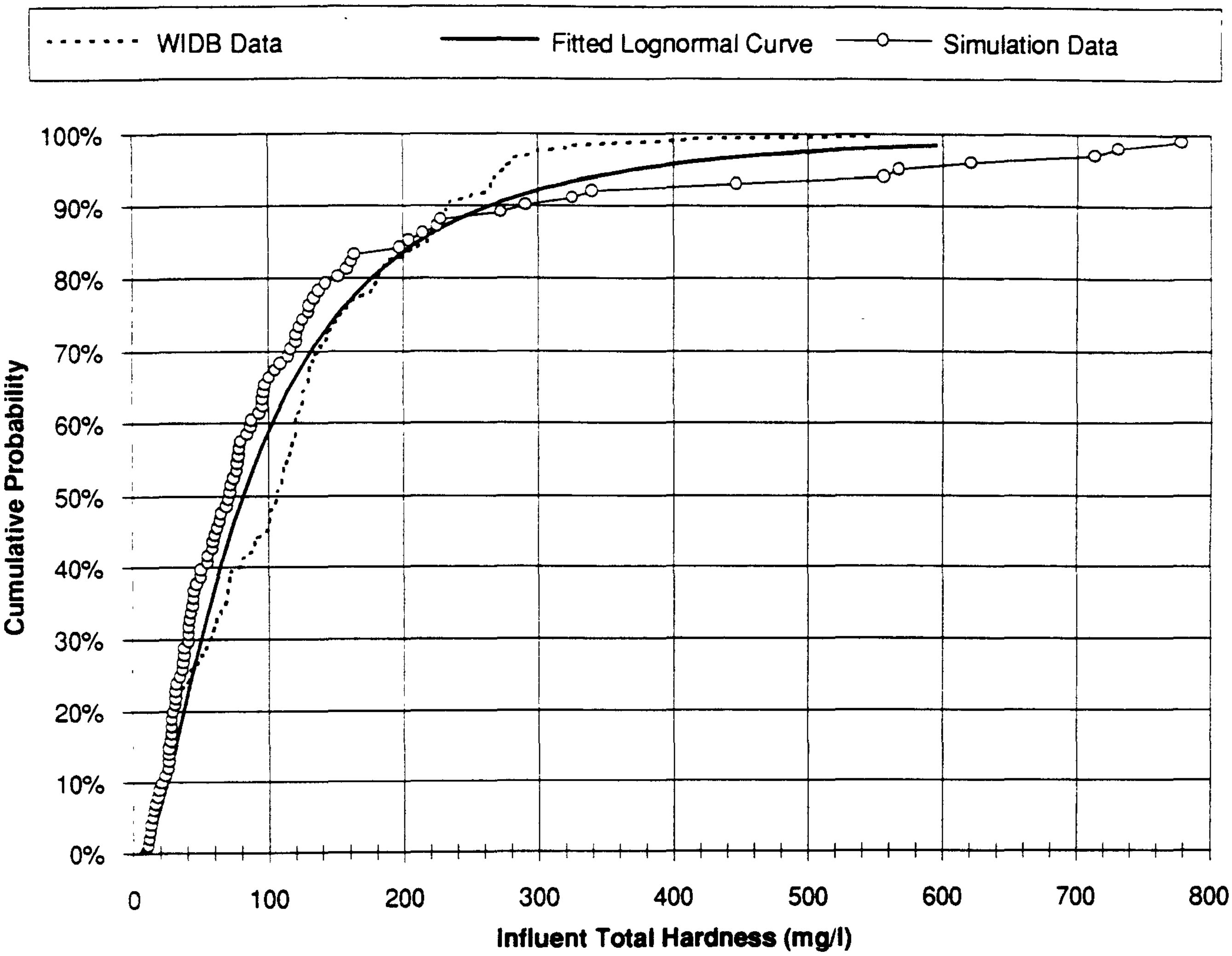
Exhibit 11 compares the cumulative distributions of the WIDB data set, the fitted cumulative lognormal distribution derived from those data, and the cumulative distribution of the 100 simulated data points. As shown there, the WIDB data and the fitted curve diverge from one another through the middle portion of the distribution, with the computed distribution favoring lower total hardness values.

3.8 Calcium Hardness

It was assumed *a priori* that calcium hardness values would be related to total hardness. However, there were only limited data available providing both total hardness and calcium hardness values for raw water samples. Using the limited information provided in a preliminary version of the recently completed lime softening survey by AWWA, its was estimated that typically calcium hardness contributes about 75% of total hardness. Therefore, the calcium hardness values were obtained for the 100 simulated data points simply by multiplying the obtained total hardness value by 0.75.

Exhibit 11

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Influent Total Hardness



3.9 Turbidity

The turbidity values for the 100 simulated SNS data points were derived from a lognormal distribution based on data from the WIDB. The summary statistics for the WIDB data and the 100 simulated points are shown below.

Summary Data for Turbidity in SNS Systems
(Concentrations in NTU)

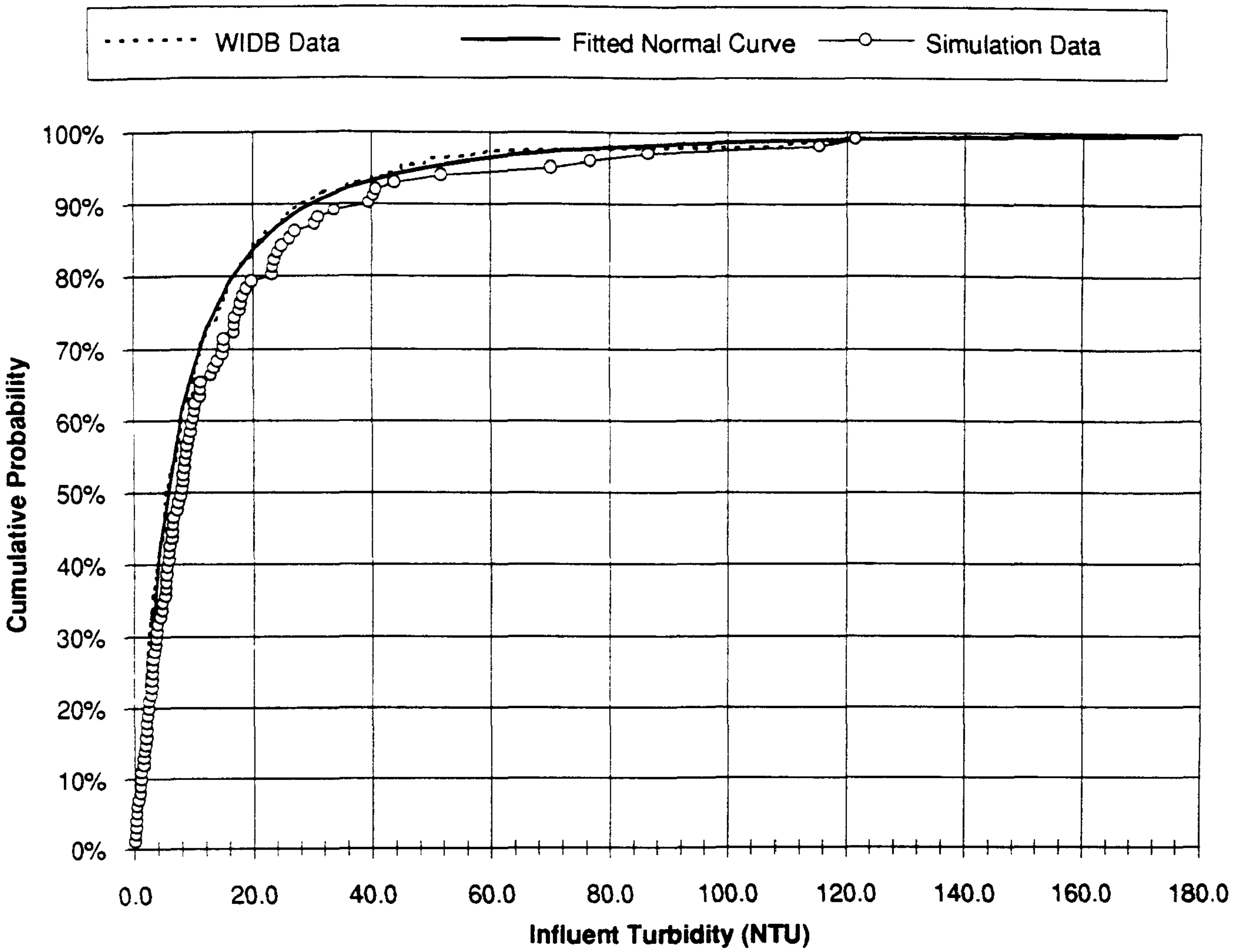
	WIDB	Simulated Data Set
N Count	300	100
Minimum Value	0.11	0.17
Maximum Value	170	121
Median	5.3	8.0
Arithmetic Mean	12.4	15.1
Standard Deviation	20.6	21.7
Log Mean	1.69	1.96
Log Standard Deviation	1.33	1.32
exp(Log Mean)	5.4	7.1

Lower and upper bounds of 0.01 and 1,000 NTU were set for the 100 simulated data points.

Exhibit 12 compares the cumulative distribution of the 300 WIDB data points, the fitted cumulative lognormal distribution, and the cumulative distribution of the 100 simulated data points selected through the Monte Carlo simulation. As evidenced from this exhibit, as well as from the summary data shown above, the underlying WIDB turbidity data and the 100 simulated turbidity data points for the SNS systems are highly similar.

Exhibit 12

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Influent Turbidity



3.10 Ammonia

The underlying ammonia data used for this analysis was 55 SNS samples from the AWWA Disinfection Survey. Shown below are the summary statistics from that data set, along with the statistics for the 100 simulated data points.

Summary Data for Ammonia in SNS Systems
(Concentrations in mg/L)

	DS	Simulated Data Set
N Count	55	100
Minimum Value	0.01 mg/L	0.003 mg/L
Maximum Value	3.2 mg/L	1.4 mg/L
Median	0.05 mg/L	0.05 mg/L
Arithmetic Mean	0.16 mg/L	0.11 mg/L
Standard Deviation	0.44 mg/L	0.17 mg/L
Log Mean	-1.22 (Base 10)	-1.23 (Base 10)
Log Standard Deviation	0.545 (Base 10)	0.465 (Base 10)
exp(Log Mean)	0.06 mg/L	0.06 mg/L

The analysis of the ammonia data from the Disinfection Survey was provided by Dr. Charles Haas of Drexel University. His analysis to determine the best distributional fit differed from that used for other data in that the log transformation for evaluating the lognormal distribution used base 10 logs rather than natural logs used for other variables. Also, the goodness of fit test used was the Kolmogorov-Smirnov test rather than the Shapiro-Wilk test. Although these differences are unlikely to have affected the outcome of this analysis, these inconsistencies will be corrected in future iterations.

The ammonia data were found to have a better fit to the lognormal distribution. Lower and upper bounds of 0 and 4 mg/L were placed on the ammonia values obtained for the 100 simulated systems.

3.11 Alkalinity

It was assumed that alkalinity levels would be related to the pH. Regression analyses were conducted for this relationship using 288 data points from the WIDB data base having both pH and alkalinity data for SNS systems. Because pH is a log-transformed value, only the following two relationships were tested:

$$Alkalinity = a + b\,pH \quad (r^2 = \mathbf{0.591}; W = \mathbf{0.956})$$

$$\ln(Alkalinity) = a + b\,pH \quad (r^2 = \mathbf{0.725}; W = \mathbf{0.983})$$

Based on both the r^2 and the W statistics, the second of these relationships was selected to characterize the relationship between alkalinity and pH. The following parameters were obtained for this regression:

Intercept (a):	-6.56
Slope (b):	1.29917
Root MSE:	0.4659

Exhibit 13 shows the relationship between the underlying WIDB data and the regression relationship derived from those data. Exhibit 14 shows the scatter of the 100 simulated data points relative to the regression line.

Exhibit 13

REGRESSION ANALYSIS
Actual Data and Regression-predicted Influent Alkalinity

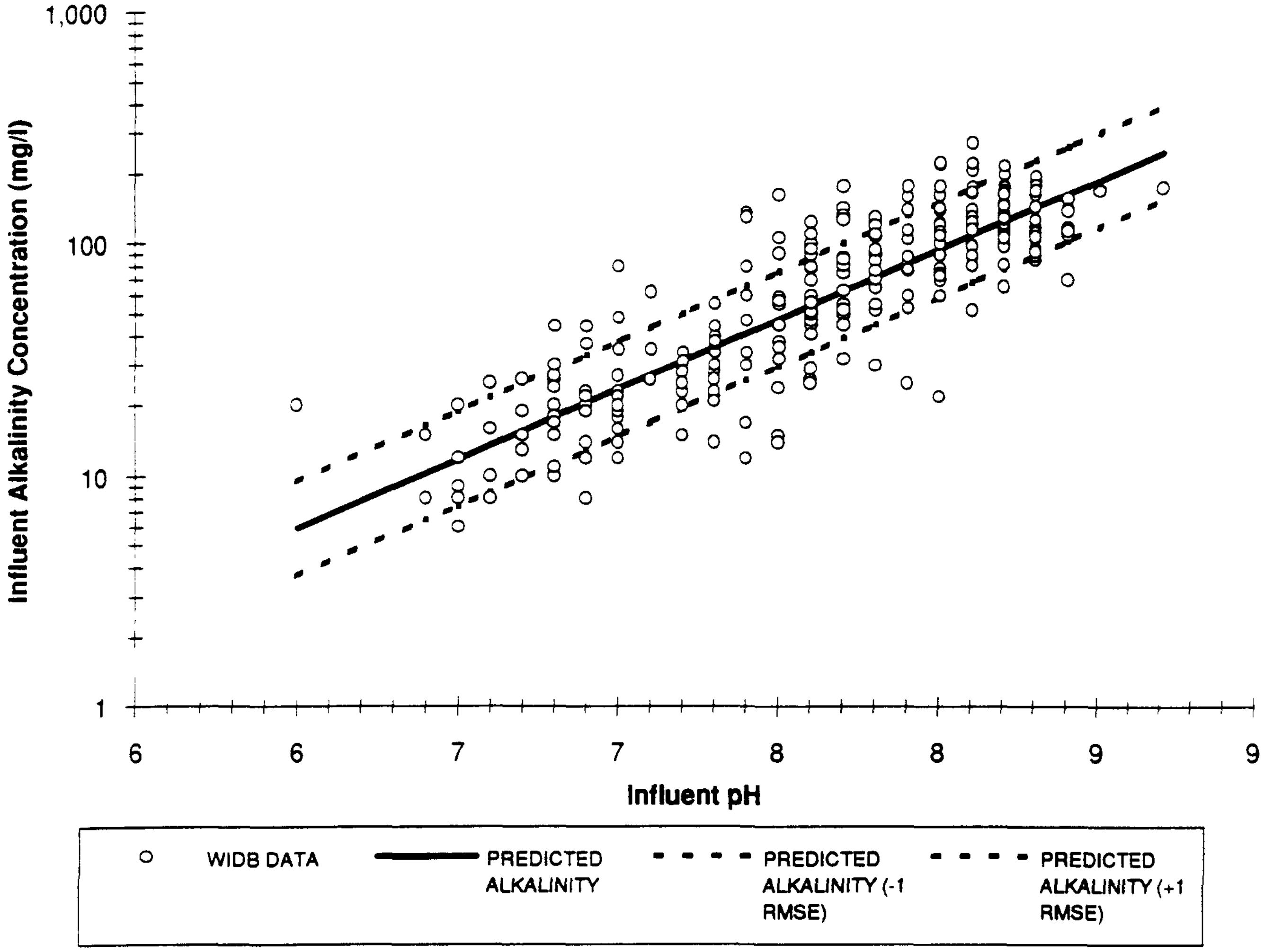
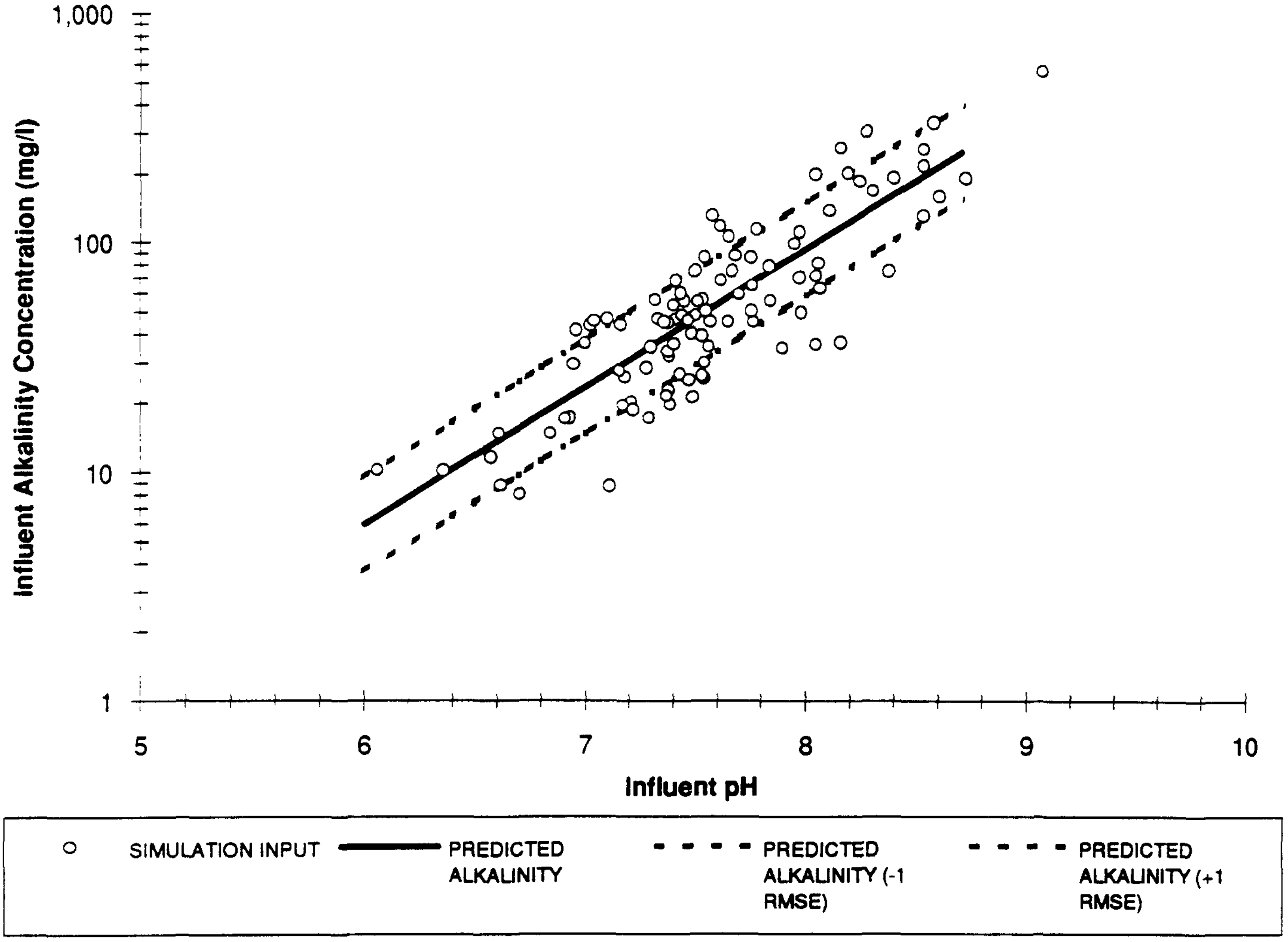


Exhibit 14

REGRESSION ANALYSIS
Simulation and Regression-predicted Influent Alkalinity



3.12 Distribution System Residence Time

The values for distribution system residence time for the 100 simulated SNS systems were obtained from a lognormal distribution derived from data provided in WIDB. Summary statistics for hose data, and the 100 simulated data points are provided below,

Summary Data for Distribution System Residence Time in SNS Systems
(Values in Days)

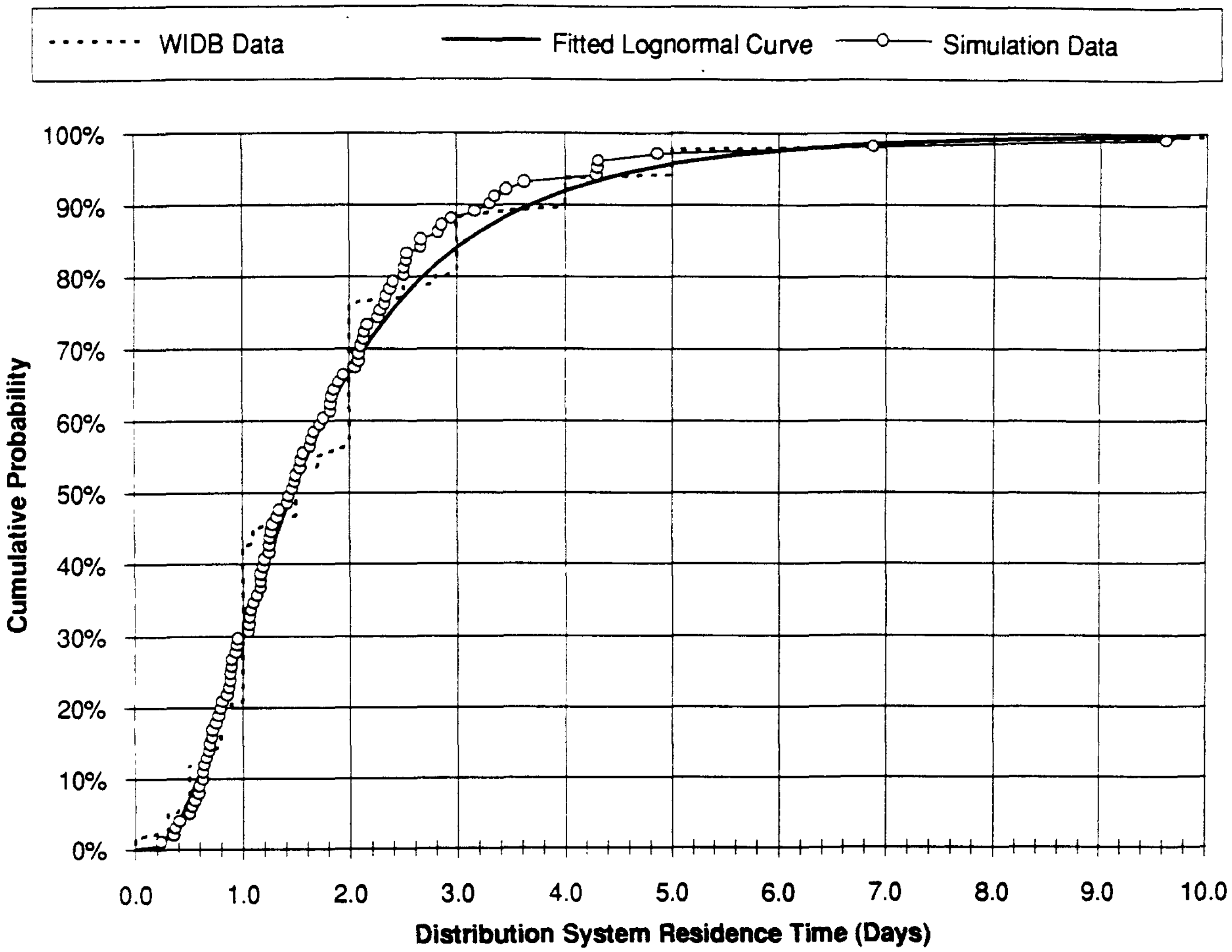
	WIDB	Simulated Data Set
N Count	204	100
Minimum Value	0.0	0.24 days
Maximum Value	10 days	9.6 days
Median	1.5 days	1.5 days
Arithmetic Mean	1.8 days	1.8 days
Standard Deviation	1.4 days	1.4 days
Log Mean	0.381	0.358
Log Standard Deviation	0.719	0.659
exp(Log Mean)	1.46 days	1.43 days

Based on the goodness of fit test performed, the lognormal distribution was selected, with the Log Mean and Log Standard Deviation from the WIDB data used as the parameters for that distribution. Lower and upper bounds of 0.1 and 10 days were used for the simulated data points obtained from this distribution.

Exhibit 15 provides a comparison of the cumulative distribution of the WIDB data, the fitted cumulative distribution derived from those data, and the cumulative distribution of the 100 simulated data points. As shown there, these data compare favorably. It should be noted that the "step" appearance of the WIDB data points relative to the derived distribution reflect the apparent tendency for some of those data to have been reported in whole day values.

Exhibit 15

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Distribution System Residence Time



3.13 Average Daily Flow

Average daily flow values for the simulated data set were based on a lognormal distribution derived from 303 WIDB data points for SNS systems. Summary statistics for those data and for the 100 simulated data points are presented below.

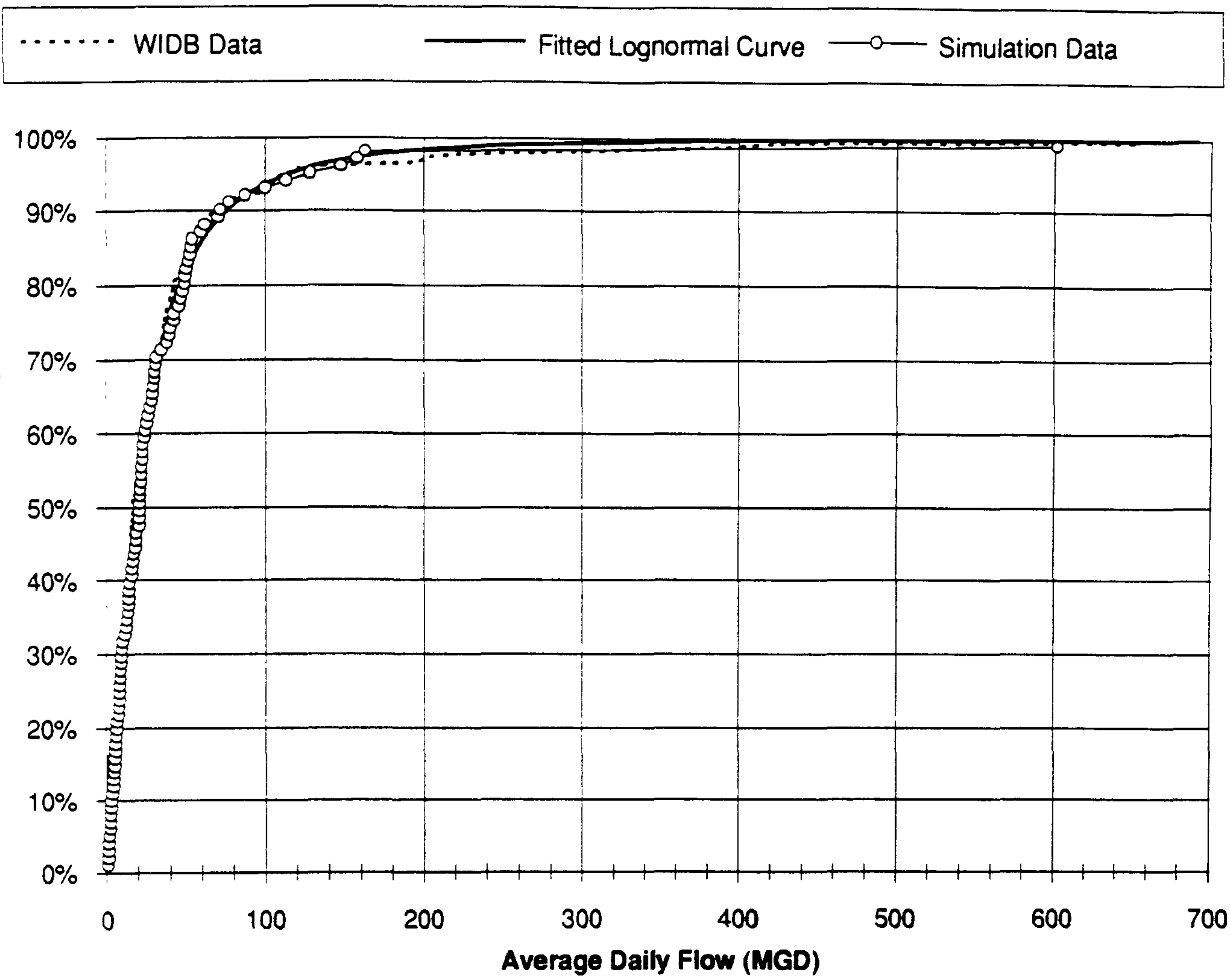
Summary Data for Average Daily Flow in SNS Systems
(Values in Millions of Gallons per Day)

	WIDB	Simulated Data Set
N Count	303	100
Minimum Value	1.0 MGD	1 MGD
Maximum Value	654 MGD	602 MGD
Median	16.5 MGD	21.0 MGD
Arithmetic Mean	35.6 MGD	36.1 MGD
Standard Deviation	65.6 MGD	66.1 MGD
Log Mean	2.85	2.91
Log Standard Deviation	1.149	1.16
exp(Log Mean)	17.3 MGD	18.4 MGD

Upper and lower bounds of 1 and 1,000 MGD were established for the simulated data set. Exhibit 16 provides a comparison of the cumulative distribution of the WIDB data, the cumulative lognormal distribution derived from those data, and the cumulative distribution of the 100 simulated data points. As shown there, the middle portion of the fitted distribution diverges slightly from the WIDB data, with the fitted distribution favoring slightly higher flow values, which are reflected in the simulated data set.

Exhibit 16

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Average Daily Flow



3.14 Lime Dose

Lime dose values for the SNS systems were based on a normal distribution derived from 123 WIDB data points. The summary statistics for these 123 data points and the simulated data points are presented below:

Summary Data for Lime Dose in SNS Systems
(Values in mg/L)

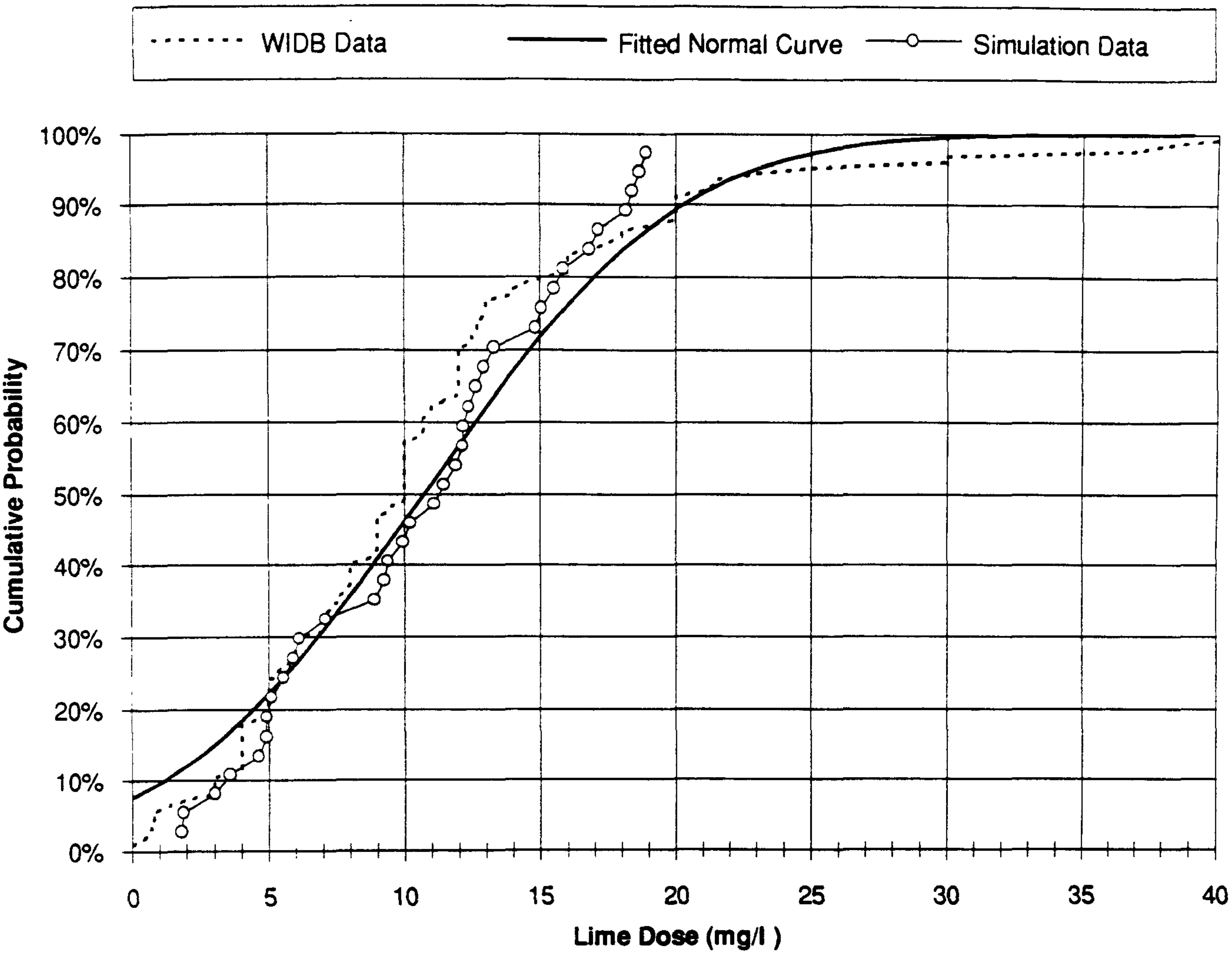
	WIDB	Simulated Data Set
N Count	123	36
Minimum Value	0.05 mg/L	1.8 mg/L
Maximum Value	40 mg/L	18.7 mg/L
Median	10 mg/L	11.1 mg/L
Arithmetic Mean	10.7 mg/L	10.4 mg/L
Standard Deviation	7.4 mg/L	5.0 mg/L
Log Mean	2.07	2.20
Log Standard Deviation	0.951	0.634
exp(Log Mean)	7.9 mg/L	9.0 mg/L

As indicated above, there were only 36 data points in the simulated data set for lime dose for the SNS systems. This reflects the information derived from WIDB showing that only 36% of the SNS systems use lime. Although lime dose values were selected for each of the 100 simulated systems, a second step was carried out where 64 of those systems were randomly selected, and the lime dose values for those were set to 0.

Exhibit 17 provides a comparison of the cumulative distribution of the 123 WIDB data points, the cumulative normal distribution derived from those data, and the cumulative distribution of the 36 non-zero simulated values. In the range of the distribution between approximately the 40th and 90th percentiles, the fitted distribution shows higher values than the underlying WIDB data.

Exhibit 17

Comparison of Cumulative Probability Distributions For Actual and Model-predicted Lime Dose



3.15 Alum Dose

The alum dose was assumed to be related to the influent TOC values, with the expectation that higher alum doses would be observed in systems having higher influent TOC levels. The relationship between alum dose and TOC was using 54 data points from SNS systems in the WIDB. All four relationships were tested, the summary statistics for which are shown below.

$$AlumDose = a + b\,TOC \quad (r^2 = 0.472; W = 0.969)$$

$$\ln(AlumDose) = a + bTOC \quad (r^2 = 0.322; W = 0.954)$$

$$AlumDose = a + b\ln(TOC) \quad (r^2 = 0.340; W = 0.937)$$

$$\ln(AlumDose) = a + b\ln(TOC) \quad (r^2 = 0.279; W = 0.956)$$

All of the r^2 values were fairly poor, and although these and the Shapiro-Wilk statistic for the AlumDose vs. TOC regressions had the best values, the final regression form chosen for both the SNS group was ln(AlumDose) vs. TOC. This deviation from the established protocol was necessitated because the use of the AlumDose vs. TOC form, owing to the overall poor fit, resulted in a substantial number of values selected that exceeded the established lower and upper bounds for alum dose. These bound were 0.5 and 300 mg/L.

The parameters obtained for this regression were as follows:

Intercept (a):	1.646
Slope (b):	0.2521
Root MSE:	0.8143

Exhibit 18 provides a plot of the WIDB data against the selected regression line. Exhibit 19 presents a similar plot for the 100 simulated data points.

It should be noted that in subsequent iterations of this analysis, it is intended that the alum dose will be tested for its relationship with turbidity to determine whether it is a better predictor of alum dose than TOC.

3.16 Prechlorination

The prechlorination treatment variable was a different form from the others. The purpose of this variable was to indicate whether a particular system does or does not practice prechlorination. Based on 313 SNS data points in the WIDB, it was determined that 82% practice prechlorination. Therefore, 82 of the 100 simulated SNS systems were selected randomly and designated as performing prechlorination.

Exhibit 18

REGRESSION ANALYSIS
Actual and Regression-predicted Alum Dose

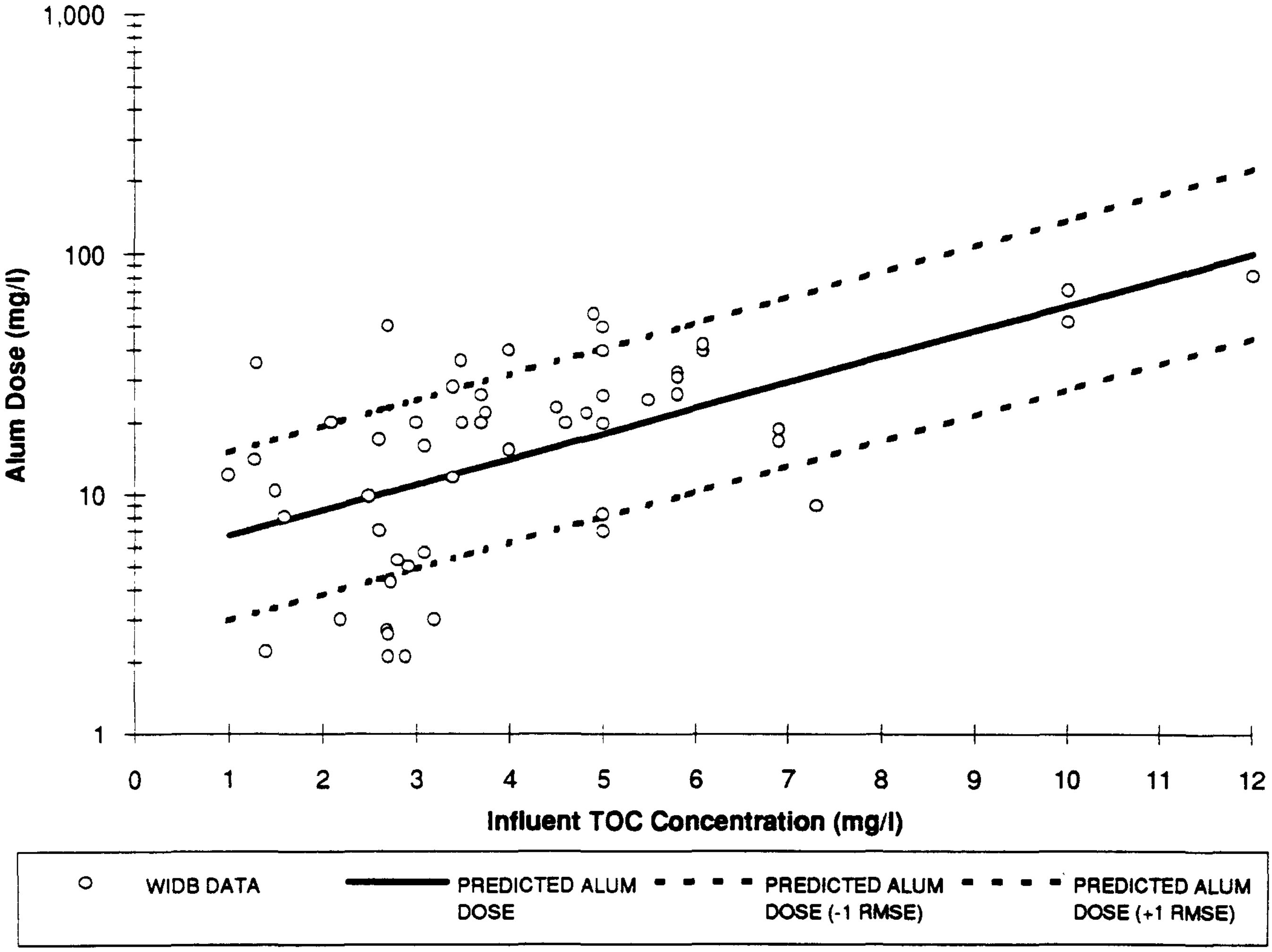
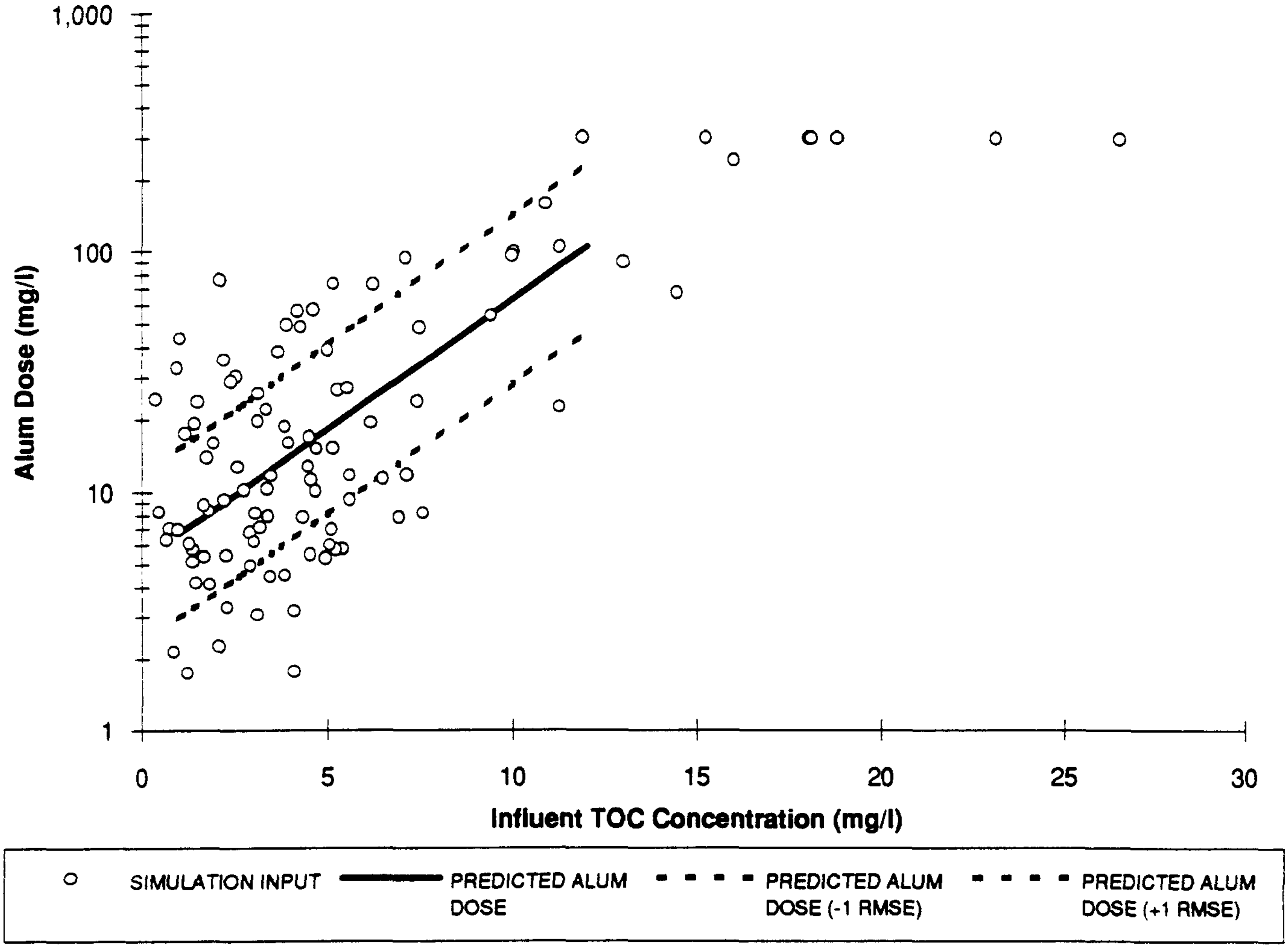


Exhibit 19

REGRESSION ANALYSIS
Simulation and Regression-predicted Alum Dose



4. DISCUSSION

In the preceding sections we have described the assumptions and procedures used to perform Monte Carlo simulations resulting in the creation of 100 simulated water supplies that are subsequently used in a model to evaluate compliance choices to meet regulatory options being considered for disinfection by-products. An obvious concern that arises with respect to considering the validity of the approach being used is that of the representativeness of the simulated data. That is, how well do these 100 simulated water supplies represent the population of large, filtering, non-softening surface water supplies with respect to the levels of the raw water quality parameters and the water treatment characteristics used to describe them? The representativeness question can be broken into several parts:

1. How representative are the underlying data set used of the real world distribution of values for each variable?

2. How well do the probability density function developed from those data represent the real world distributions?

3. How well does the simulated set of values reflect the probability density function from which it is obtained?

Most of the variables included in the simulated data set were based on 200 to 300 data points provided by the WIDB. This is a substantial number of data points, and comprises a large proportion of the total number of these types of supplies in the nation. It should be noted that the data in the WIDB was provided by the water supply, and was not an independent sampling and analysis. Consequently, there were no uniform QA/QC procedures applied to ensure that all of the data reported was of a comparable level of reliability. Nevertheless, it is assumed that, overall, the WIDB data are reliable and provide a reasonable representation of the SNS supplies.

For some of the variables, the number of data points being relied upon to describe the national distribution of values were, however, much lower than the 200 to 300 noted above. Three of these variables are of particular concern: TOC, UV-254 and bromide. The TOC analysis was based on 84 WIDB data points. While the range of values provided by these 84 data points appear to be reasonably consistent with the levels of TOC expected in the raw water for these types of supplies, it has been conjectured that these values may have a slight upward bias in these data. Since TOC is not a water quality measure that is normally required of water supplies, the limited number of systems that had the data available to report on the WIDB questionairre may be systems that have a TOC-related problem. The UV-254 relationship with TOC was based on a very small data set, 23 observations from the JMM study. The bromide data was based on only 18 data points from the JMM study, and there is some concern that because the JMM data set was heavily represented by California water systems, these data may have an upward bias. However, this may not be that significant because the brominated THM with the highest potency factor, bromodichloromethane, is the lest affected of all the brominated species, by bromide levels. More relevant may be the predicted brominated HAAs when potency factors become available for these compounds.

Another major concern is the data used to characterize prechlorination. Data from the WIDB did not distinguish the precise point of prechlorination preceeding filtration. the WTP model (described by Gelderloos et al., 1992) assumed that all points of prechlorination occurred prior to the rapid mix (that is 82% of the SNS systems applied chlorine there). Since a significant number of systems are known to prechlorinate just prior to filtration, the WTP model assumption over-represents the status quo conditions for predicting DBP formation resulting from prechlorination. Future model runs will attempt to use more representative prechlorination data if it becomes available.

The WTP model assumption of 82% prechlorinating at the rapid mix probably more closely represents industry practice prior to the 1979 TTHM regulation. If this is true, then the WTP model can be crudely tested for its reliability for predicting treatment changes to meet a TTHM standard of 100 µg/L with survey results indicating actual treatment changes made by the industry to comply with the TTHM standard. This is discussed further in the paper by Cromwell et al. (1992) and Gelderloos et al. (1992).

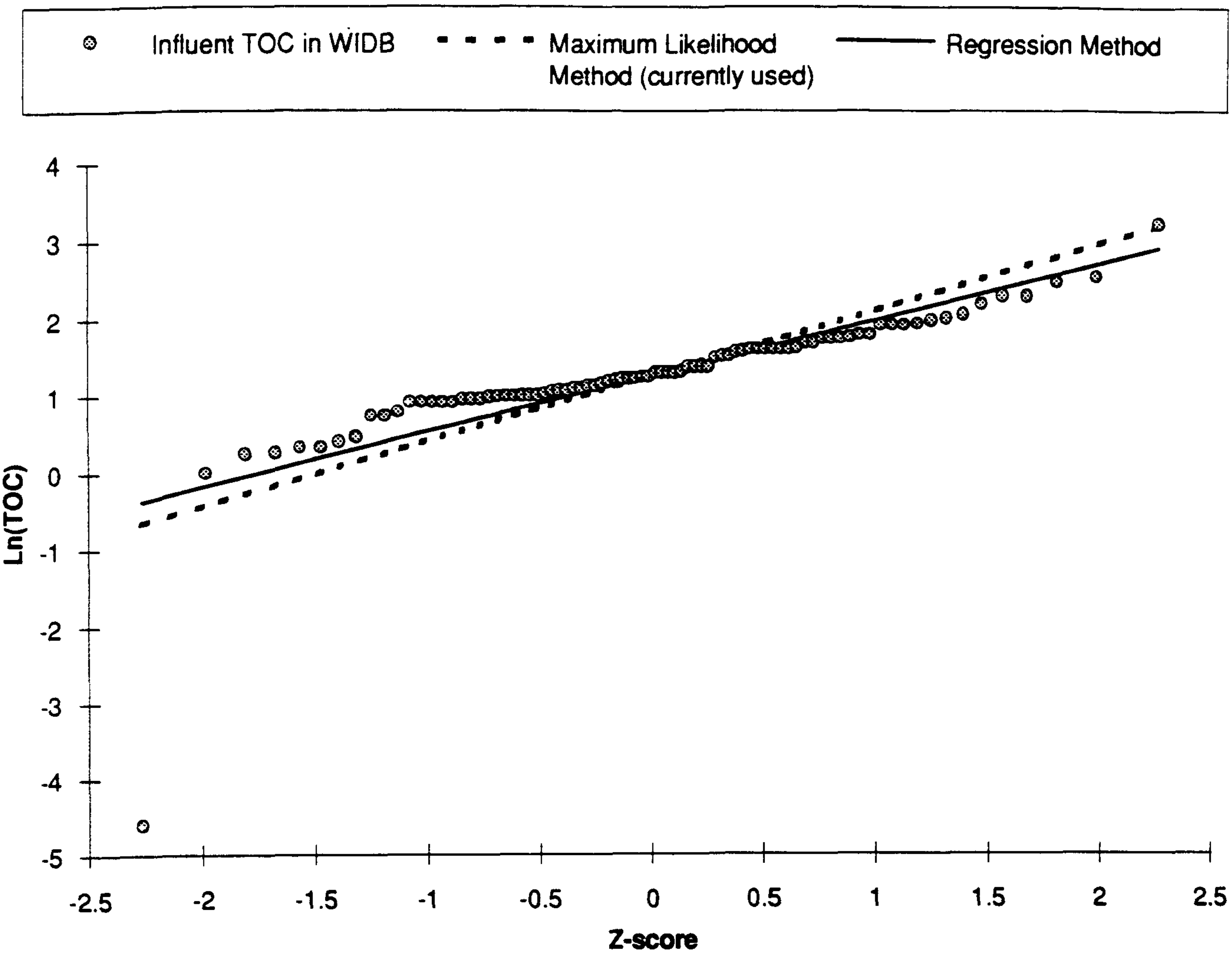
The particular concern about the representativeness of these three variables relates to their critical role in predicting by-product levels. Referring back to the TTHM formation equation presented in Section 2, all three of these variables are incorporated directly. An upward bias in TOC, coupled with the dependency of UV-254 on those TOC values, could result in an overestimate of TTHM formation. An upward bias in the bromide levels used would not have a very large impact on the TTHM estimates, since the exponent for bromide in the formation equation is very small (0.036). However, the bromide level does have a significant effect on the distribution of the individual THM species, with higher bromide levels favoring the more highly brominated species, which also have higher carcinogenicity potency factors .

The question of the how well the probability distributions derived from the available data correspond to the underlying data is also critical As noted previously, the selection of the normal and lognormal distributions to describe the data was made *a priori*.. While these distribution forms have been found to correspond well with environmental measurement data across most of the range of those data, they are ultimately approximations of the true distributions. Also, the procedures for estimating the parameters of those distributions can affect the degree to which they correspond to the underlying data. As noted in Section 2, the parameter estimates were made by simply using the mean and standard deviation (or log mean and log standard deviation) of the underlying data. As also noted there, an alternative approach using regression techniques has been found to provide better estimates of distributional parameters for environmental data.

An example of the difference between these is shown in Exhibit 20 for TOC. This exhibit provides a plot of the order statistics of the underlying WIDB data, a plot of the distribution used in this analysis, and a plot of the distribution of TOC based on the regression method for estimating the parameters. As can be seen there, the regression method appears to provide a better fit to the underlying data than the method used here to estimate the parameters. This is mainly due to the effect of the two extreme values (especially the one extremely low value) on the estimation of the σ parameter, which corresponds to the slope of the lines presented in this exhibit. Using the current method, these values have more weight in estimating this parameter than they do with the regression method. (It is also reasonable to consider eliminating these extremes, which would further improve the fit to the WIDB data.) In subsequent iterations of this analysis, the regression method will be more fully explored for making the parameter estimates.

Exhibit 20

Comparison of TOC Distributions to Underlying Data Using Alternative Methods for Estimating Parameters



The third question, that of the adequacy of 100 data points to represent the universe of SNS water supplies, has two aspects to it. The first is the representativeness of those data points with respect to any one individual parameter. It can be shown from basic sampling theory that 100 samples are more than adequate to estimate the mean and variance of a population. Generally, it can be shown that beyond approximately 30 samples for a population the size of the SNS systems, there is minimal increase in the precision of the estimates of the mean and variance by taking additional samples. A smaller sample size will, however, tend to provide less reliable information about the extreme values (for example, the 90th or 95th percentile value).

The second aspect of the question about the adequacy of the 100 simulated systems relates to the representativeness of the combinations of values for all variables. As with the representativeness of individual variables, the 100 simulated sets are expected to be reasonably representative of the range of combined values in terms of their effect on TTHM formation. It is expected, however, that the 100 simulated systems do not provide fully representative information on the systems that may have simultaneously occurring extreme values of several parameters that have a significant direct effect on by-product formation. Subsequent iterations of this analysis will explore these questions in more detail for inclusion in a more fully developed uncertainty assessment.

5. REFERENCES

Amy, GL; Chadik, PA; and Chowdhury, ZK. (1987) *Developing Models for Predicting Trihalomethane Formation Potential and Kinettics.* Jour. Amer. Water Works Assoc., Vol. 79, No. 7, Pg. 89.

Cromwell, J; Xhang, X; Letkiewicz, F; Regli, S; and Macler, B. (1992) *Preliminary Regulatory Impact Analysis: Trade-offs in Regulation of Disinfection By-Products.* Unpublished.

Gelderloos, AB; Harrington, GW; Owen, DM; Regli, S; Schaefer, JK; Cromwell, JE; and Xhang, X. (1992) *Simulation of Compliance Choices for the Regulatory Impact Analysis.* Unpublished.

Harrington, GW; Chowdhury, ZK; and Owen, DM. (1991) *Integrated Water Treatment Plant Model: A Computer Model to Simulate Organics Removal and DBP Formation.* Proceedings of the 1991 AWWA Annual Conference, Philadelphia, PA.

Helsel, DR. (1990) *Less than Obvious. Statistical Treatment of Data Below the Detection Limit.* Environmental Science and Technology, Vol. 24, No. 7, pp. 1766-1774.

Helsel, DR and Gilliom, RJ. (1988). *Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 2. Verification and Applications.* Water Resources Research, Vol. 24, No. 12, pp. 1997-2004.

Patania, NL. (1991). *AWWA D/DBP Database and Model Project.* Contract No. 1-90. Final Report addressed to Edward Means. August 26, 1991.

Travis, CC and Land, ML. (1990) *Estimating the Mean of Data Sets with Non-Detectable Values.* Environmental Science and Technology, Vol. 24, No. 7, pp. 961-962.