APPENDIX A

Characterization of Uncertainty in Estimates

CONTENTS

ABBREVIATIONS AND ACRONYMS v

A.1.0 INTRODUCTION 1

A.2.0 CHARACTERIZING UNCERTAINTY 5
A.2.1 PROBABILISTIC PARAMETERS 6
A.2.1.1 Covariance of Functions 8
A.2.1.2 Linear Functions of Uncertain Variables 9
A.2.2 USE OF LOGNORMAL PROBABILITY DENSITY FUNCTIONS 13
A.2.2.1 Lognormal Probability Functions and Parameters 15
A.2.2.2 Lognormal Parameter Estimates From Statistical Data Analysis 16
A.2.2.3 Probabilistic Parameters for Products of Lognormal Variables 20
A.2.2.4 Probabilistic Parameters for Quotients of Lognormal Variables 22
A.2.2.5 Probabilistic Parameters for Power Series of Uncertain Variables 23
A.2.2.6 Probabilistic Parameters for a General Function of a Lognormal Variable 25

A.3.0 PROBABILISTIC ESTIMATES 27
A.3.1 PROBABILITY TRUE VALUE OF X DOES NOT EXCEED GOAL G 27
A.3.2 NON-EXCEEDANCE ESTIMATES OF X, X_{PN} 28
A.3.3 ESTIMATE OF X REQUIRED TO MEET GOAL G WITH SPECIFIED PROBABILITY 30

A.4.0 BAYESIAN ESTIMATES OF EXPECTED VALUES AND COEFFICIENTS OF VARIATION 31

A.5.0 REFERENCES 35

FIGURES

A-1 Probability (true) X < Goal G, P[X<G] v. Ratio of Expected Value of X to G, E[X]/G39.....39
A-2 Non-Exceedance Ratio X_{Pn}/E[X] v. Non-Exceedance Probability, Pn 40
A-3 E[X] Required to Meet Goal G, E[X_G], with Specified Probability Pr=P[X<G] 41
A-4 Ratio UB[X]/E[X] and LB[X]/E[X] 42

ABBREVIATIONS AND ACRONYMS

BHSS	Bunker Hill Superfund Site
CDF	cumulative density function
CDR	Coeur d'Alene River
cfs	cubic foot per second
cy	cubic yard
Eq	equation
FP	flood plain sediment
FS	feasibility study
HELP	hydrologic evaluation of landfill performance
LB	lower bound
lb/d	pound per day
LCL	lower confidence limit
M/T	metal mass per unit time
M/V	metal mass per volume of water
µg/L	microgram per liter
mg/kg	milligram per kilogram
PAT	probabilistic analysis tool
PDF	probability density function
ppm	part per million
PRB	permeable reactive barrier
RI	remedial investigation
RI/FS	remedial investigation/feasibility study
RLP	relative loading potential
SFCDR	South Fork Coeur d'Alene River
TCD	typical conceptual design
tons/cy	tons per cubic yard
UB	upper bound
UCL	upper confidence limit
USGS	United States Geological Survey
V/T	volume per unit time
yr	year

A.1.0 INTRODUCTION

This appendix provides a general discussion of uncertainty associated with quantitative estimates of physical variables of interest. For the probabilistic analysis of metal loadings, these variables include stream discharges, concentrations, loadings, ratios of loading or concentration to goals (such as AWQC) or other targets (e.g., TMDL loading capacities), as well as source volumes, relative loading potentials, remediation factors, and the like. The discussion introduces general concepts used to characterize and quantify uncertainty as part of the probabilistic analysis. The discussion centers on the need to quantitatively *estimate* physical variables of interest under conditions of uncertainty. The estimates may relate to current conditions or *predictions* of potential future conditions.

In practice, physical variables of interest must be estimated by some method. Depending on the situation, estimation methods may be direct or indirect measurements, calculations using mathematical models (which may be simple or complex), inferences or interpretations, hypotheses, assumptions, or some combination of methods. These methods, which generally include "objective" and "subjective" aspects, are never truly exact.

All estimates of physical variables are approximations having some degree of potential inaccuracy, or "error," relative to exact, or *true*, values. Where the potential estimation error is significant, the variables may be characterized as "uncertain variables."[1] Very simply:

> The exact or true value of an uncertain variable cannot be estimated or predicted without potential estimation inaccuracy or "error."

To deal with this potential error, uncertain variables are treated probabilistically to make probabilistic estimates of their inherently uncertain true values. Equivalent terms used in this document (and generally in the literature) for uncertainty and related inaccuracy may include, depending on context, "error," "estimation error," or "estimation uncertainty."[2]

---

[1] "Random variables" is the technically correct term (from probability theory). However, the term "random variables" can be misleading since it could be erroneously construed to mean that such variables can take on any value with equal probability—which is not the general case. Probability theory allows quantitative predictions of "random" phenomena.

[2] The inaccuracy more precisely includes inexactitude, as probability bounds (including exceedance/non-exceedance estimates, confidence and tolerance bounds, etc.) can provide estimates that are accurate but not necessarily precise, since increasing accuracy usually comes at the cost of decreased precision for a given state of information or data. Unfortunately, the semantics can get somewhat confusing and tortured.

Uncertainty stems from imperfect information and incomplete knowledge. Only with perfect information and complete knowledge is there no uncertainty or potential error. For the case of metal loadings and concentrations, estimation uncertainty comes from several interrelated sources. These uncertainty sources include:

- Limited available measurement and observational data
- Simplifications and approximations used in estimating
- Natural variability including natural complexity, heterogeneity, and variability in space and over time
- Limited understanding of how the pertinent natural systems interact and behave over space and time
- Incomplete knowledge of how hypothetical or actual remedial actions will perform over time

At least three types of estimation uncertainty exist: natural variability, model uncertainty, and parameter uncertainty. Natural variability over space or time includes the intrinsic complexity of the physical phenomenon of interest. Model uncertainty comes from inaccuracy in the deterministic or probabilistic models used to characterize the natural variability of the phenomenon of interest. Parameter uncertainty (also known as statistical uncertainty) comes from inaccuracy in the parameters of the deterministic and probability models. These three types of uncertainty overlap to some extent.

Natural variability is the combination of two effects: (1) the practically irreducible uncertainty due to our limited quantitative and predictive knowledge of the fundamental physical mechanisms and interactions underlying the phenomenon of interest, and (2) the fundamentally probabilistic nature of the phenomenon itself. In principle, advancements in fundamental knowledge could reduce the first effect, at a cost, but not the second. From a practical standpoint, natural variability can be considered "intrinsic, fundamental, irreducible" uncertainty, reflecting the inexactitude of available knowledge.

From a predictive standpoint, model uncertainty reflects the fact that real-world data do not *exactly and perfectly* fit any theoretical probability distribution (this is true for both parametric and non-parametric or empirical distributions—e.g., see Hanover 1980, p. 367). However, although the fits are always approximations, they may be quite adequate for practical purposes. Parameter uncertainty, reflecting inaccuracy in the model parameters, can be virtually eliminated with sufficient data, at least in theory.

The combination of model and parameter uncertainty means that probabilities are themselves estimates, and not exact specifications of the inexactitude of available knowledge. In particular, probabilities are statements of knowledge that are always conditional on available information and its interpretation. This fundamental fact is implicit in all the probabilities estimated or discussed in this document. Since interpretation is required, so is understanding and sound judgment. Keeping these limitations in mind, for practical purposes, probabilistic estimates can provide a useful, quantitative specification of potential outcomes.

Probabilistic estimates use the concepts and mathematics of probability theory, including random variables and random processes ("laws of chance") and error propagation. Although the fundamental basis of these concepts and mathematics is well established, their practical application depends on context. Background for the approach used to make probabilistic estimates is summarized in the following sections.

A.2.0 CHARACTERIZING UNCERTAINTY

This section introduces the basic concepts of probabilistic parameters and probability density functions (PDFs). Probabilistic parameters and PDFs are used to quantitatively characterize uncertainty and make probabilistic estimates.

To simplify presentation and discussion, the symbol X is used to generally represent any uncertain variable such as metal loadings, source volumes, or remediation factors. Also, the uncertain variables may or may not be functions of other variables, which may or may not be treated probabilistically.

Probabilistic parameters are similar to statistical measures, such as averages, standard deviations, variances, and correlation coefficients. However—and this is key—probabilistic parameters are *not limited* to measurement data. They can be used to characterize uncertainty, including natural variability or potential estimation error, in any kind of estimate—whether it is measured, calculated, inferred, hypothesized, opined, or assumed.

The PDF of X, PDF_X, defines the probability (density) of the true (but uncertain) value X over its full range of potential values. PDF_X also determines the cumulative density function (CDF) of X, CDF_X. Given PDF_X, CDF_X defines the probability that any particular *estimate* of X will not be exceeded by the *true* (but uncertain) value of X. Where X is restricted to positive values, CDF_X is the integral of PDF_X from zero to particular estimates of X over the range of possible values.

PDF_X is necessary to make the most useful probabilistic estimates of the true, but uncertain, value of X. These estimates include:

- the *probability* that X exceeds a given estimate
- the *estimate* that is not exceeded by X with a specified probability (e.g., 95 percent)
- the *estimate* required to meet a given value of X with some specified probability

As used here for characterizing the uncertainty in a variable, probabilities, and thus PDFs, are used and interpreted in a Bayesian sense, as discussed in Section A.4. Therefore, over the range of potential values of a variable, the PDF weights each potential value (or estimate) by the relative degree of information or confidence (or "rational degree of belief") that the potential value (or estimate) is in fact the *true* value, which is inherently uncertain.

Lognormal PDFs are used here, as will be discussed. Examples of other PDFs commonly used for characterizing uncertainty include normal, uniform, Poisson, and beta distributions (e.g., Benjamin and Cornell 1970, Harr 1987). Note that a variable's PDF implicitly includes its probabilistic parameters.

The probabilistic parameters of X can be analyzed independent of the PDF of X. Furthermore, where X is a function of other (input) variables, the probabilistic parameters of X can be mathematically determined or approximated from the parameters of the input variables without knowledge of the PDFs of the variables.

A.2.1 PROBABILISTIC PARAMETERS

Implicit in the following discussion is that the variables of interest are uncertain spatial and temporal averages. Uncertain variables are characterized by their probabilistic parameters, which, for the purpose of this document, include:

- Expected value of X, E[X]
- Coefficient of variation of X, CV[X]

The expected value can be considered a "best estimate" of the true, but uncertain, value. Mathematically, the expected value is an average or mean value.[3] The coefficient of variation is a measure of variability or uncertainty of potential estimates around the expected value. The expected value and coefficient of variation establish the minimum necessary information to provide a probabilistic description of a variable.

The coefficient of variation is related to the expected value through the variance, V[X]. Mathematically, the variance is the average squared deviation of the range of potential estimate from the expected value. The variance is the same as the square of the standard deviation, SD[X]. Specifically, the coefficient of variation, CV[X], is defined from the variance and expected value as follows:

$$\begin{aligned} CV[X] &= V[X]^{1/2} / E[X] \\ &= SD[X] / E[X] \end{aligned} \qquad (1)$$

[3] Since the physical variables of interest are spatial-temporal averages, the expected value is a multi-dimensional average, or simply, an "average of an average." Vanmarcke (1983) discusses effects of correlation on variability.

where:

$V[X]$	=	Variance of X
	=	$(E[X]CV[X])^2$
$SD[X]$	=	Standard deviation of X
	=	$V[X]^{1/2}$
	=	$E[X]CV[X$

For illustration, assume the following. The average and standard deviation of chemical mass loading from a given source (for given conditions and available data) are 10 lb/d and 5 lb/d, respectively (this might be reported as 10 $\pm$ 5 lb/d). This loading would be characterized as having an expected value of 10 lb/d and a coefficient of variation of 0.50 (5 lb/d divided by 10 lb/d). Also, from the definition of the coefficient of variation, note the following:

$$E[X] \pm SD[X] = E[X]\,(1 \pm CV[X])$$

Where the expected value and coefficient of variation are dependent on multiple variables, it is necessary to consider potential correlation between the variables. Correlation is quantified through the covariance or correlation coefficient. For any two variables X_1 and X_2, the covariance is $Cov[X_1,X_2]$ and the correlation coefficient is $p_{X1,X2}$. The covariance, $Cov[X_1,X_2]$, and correlation coefficient, $p_{X1,X2}$, are related through the expected values and coefficients of variation of X_1 and X_2 as follows:

$$Cov[X_1,X_2] = p_{X1,X2}E[X_1]E[X_2]CV[X_1]CV[X_2] = p_{X1,X2}(V[X_1]V[X_2])^{1/2} \quad (2a)$$

$$p_{X1,X2} = Cov[X_1,X_2] / E[X_1]E[X_2]CV[X_1]CV[X_2] = Cov[X_1,X_2] / (V[X_1]V[X_2])^{1/2} \quad (2b)$$

The correlation coefficient measures the degree of linear correlation between X_1 and X_2, and may range between plus and minus one ($-1 \leq p_{X1,X2} \leq 1$). Independent variables are uncorrelated and (always) have a correlation coefficient of zero ($p_{X1,X2}=0$). The converse, that variables with a zero correlation coefficient are independent, is not generally true. Given a pair of variables, a correlation coefficient of zero implies no linear correlation; a correlation coefficient of 1.0 implies perfect direct linear correlation (i.e., X_1 and X_2 are directly proportional); a correlation

coefficient of -1.0 implies perfect inverse linear correlation (i.e., X_1 and X_2 are inversely proportional).

In regression analysis, the square of the sample correlation coefficient, r^2, measures the strength of the relationship between x_1 and x_2 in the regression $x_1=mx_2+b$, where x_1 and x_2 represent sample measurement data of X_1 and X_2. The regression parameters m and b are estimated by the method of least squares (Natrella 1966). In this context, r^2 may be interpreted as the proportion of total variability measured by the regression. The square root of r^2 is an estimate of the correlation coefficient $p_{X1,X2}$, as used in the probability equations.

Correlation may be deterministic (i.e., functional or causal), probabilistic (i.e., stochastic or statistical), or a mix of both. Correlation occurs naturally between variables having the same or related causes. Correlation also occurs between variables that are functions of common variables.[4] The following section provides an approximation technique for estimating the covariance between functions having common uncertain variables.

A.2.1.1 Covariance of Functions

Functions of uncertain variables will be correlated whether or not the uncertain variables are themselves correlated. In general, for any two variables, say Y and Z, that are functions of common uncertain variables X_i, such that $Y=f(X_1, X_2, \ldots, X_n)$ and $Z=g(X_1, X_2, \ldots, X_n)$, will be correlated, even if the X_i are themselves independent. The covariance between Y and Z can be approximated using the following relationship based on Taylor Series expansion techniques (e.g., Benjamin and Cornell 1970, Meyer 1975, Hahn and Shaprio 1967, and Kaplan 1984); that is:

$$\mathrm{Cov}[Y,Z] = \mathbf{E}_{i=1,n}\mathbf{E}_{j=1,n} \{MY/MX_i\}_m\{MZ/MX_j\}_m \mathrm{Cov}[X_i,X_j] \qquad (3)$$

The partial derivatives in Eq 3 are evaluated at the mean, m, or expected value, $E[X_i]$, of the variables. For the case that all X_i are independent, Eq 3 reduces to the following:

$$\mathrm{Cov}[Y,Z] = \mathbf{E}_{i=1,n}\{MY/MX_i\}_m\{MZ/MX_i\}_m V[X_i] \qquad (4)$$

[4] This effect is manifested in so-called "spurious correlation" between independent variables that are multiplied or divided by a common variable (e.g., sediment load vs. flow, where sediment load is the product of flow and sediment concentration) (Chang 1998, Benjamin and Cornell 1970).

The correlation coefficient between functions Y and Z, $p_{Y,Z}$, is calculated from the Cov[Y,Z] using Eq 2 in Section A.2.1; that is:

$$p_{Y,Z} = Cov[Y,Z] / E[Y]E[Z]CV[Y]CV[Z] \quad (5)$$
$$= Cov[Y,Z] / (V[Y]V[Z])^{1/2}$$

Estimates for the expected values and coefficients of variation of functions are discussed in the following sections. Note that in the remaining sections X is used to represent any variable (including Y and Z) that is a function of other (input) variables, X_i.

A.2.1.2 Linear Functions of Uncertain Variables

Sums or differences of uncertain variables are important. A variable X that is the linear combination of "n" other variables X_i (where i=1 to n) is defined as follows where a_i represent coefficients that are not uncertain:

$$X = \Sigma_{i=1,n}\, a_i X_i \quad (6)$$

Coefficient a_i is positive for sums (addition or positive X_i) and negative for differences (subtraction or negative X_i). The following probabilistic parameters for X are based on the properties of linear functions of random variables and are mathematically exact (Benjamin and Cornell 1970).

$$E[X] = E[\Sigma_{i=1,n}\, a_i X_i] \quad (7)$$
$$= \Sigma_{i=1,n}\, a_i E[X_i]$$

$$CV[X] = V[X]^{½} / E[X] \quad (8)$$

$$V[X] = V[\Sigma_{i=1,n}\, a_i X_i]$$
$$= \Sigma_{i=1,n}\, a_i^2 V[X_i] + 2\Sigma_{i=1,n}\Sigma_{j=i+1} a_i a_j Cov[X_i,X_j]$$
$$= \Sigma_{i=1,n}(a_i CV[X_i]E[X_i])^2 +$$
$$+ 2\Sigma_{i=1,n}\Sigma_{j=i+1} p_{i,j} a_i a_j E[X_i]E[X_j]CV[X_i]CV[X_j]$$

where $p_{i,j}$ is the correlation coefficient between X_i and X_j. The sign of a given X_i, either positive or negative, affects both E[X] and CV[X]. Note the following special cases for the coefficients a_i:

- All X_i are positive and sum (i.e., $X= X_1+X_2+\ldots+X_n$): all $a_i=1$
- X_j is subtracted from X_i (i.e., $X= X_i-X_j$): $a_i=1$ and $a_j=-1$

Two particularly useful special cases are presented here: the sum of two variables and the difference of two variables.

$$X = X_1 + X_2 \tag{9}$$

$$\begin{aligned} E[X] &= E[X_1 + X_2] \\ &= E[X_1] + E[X_2] \end{aligned} \tag{10}$$

$$\begin{aligned} V[X] &= V[X_1 + X_2] \\ &= V[X_1] + V[X_2] + 2Cov[X_1,X_2] \\ &= (CV[X_1]E[X_1])^2 + (CV[X_2]E[X_2])^2 + \\ &\qquad + 2p_{i,j}E[X_1]E[X_2]CV[X_1]CV[X_2] \end{aligned} \tag{11}$$

$$X = X_1 - X_2 \tag{12}$$

$$\begin{aligned} E[X] &= E[X_1 - X_2] \\ &= E[X_1] - E[X_2] \end{aligned} \tag{13}$$

$$\begin{aligned} V[X] &= V[X_1 - X_2] \\ &= (1^2)V[X_1] + (-1^2)V[X_2] + 2(1)(-1)Cov[X_1,X_2] \\ &= V[X_1] + V[X_2] - 2Cov[X_1,X_2] \\ &= (CV[X_1]E[X_1])^2 + (CV[X_2]E[X_2])^2 - \\ &\qquad - 2p_{i,j}E[X_1]E[X_2]CV[X_1]CV[X_2] \end{aligned} \tag{14}$$

Note that for the sum, $a_1=a_2=1$ and for the difference, $a_1=1$ and $a_2=-1$. For positive correlation between X_1 and X_2 ($p_{1,2}>0$), the variance increases for the sum but decreases for the difference; and vice versa for negative correlation ($p_{1,2}<0$).

Also, for the special case that all $p_{i,j}$ are positive and equal to p, V[X] by Eq 8 reduces to the following:

$$\begin{aligned} V[X] &= \Sigma_{i=1,n}(CV[X_i]E[X_i])^2 + 2p\Sigma_{i=1,n}\Sigma_{j=i+1}E[X_i]E[X_j]CV[X_i]CV[X_j] \\ &= V[X]_{p=0} + p(V[X]_{p=1} - V[X]_{p=0}) \end{aligned} \quad (15)$$

where:

$V[X]_{p=1}$ is the upper bound of $V[X] = (\Sigma_{i=1,n}CV[X_i]E[X_i])^2 = (\Sigma_{i=1,n}SD[X_i])^2$

$V[X]_{p=0}$ is the lower bound of $V[X] = \Sigma_{i=1,n}(CV[X_i]E[X_i])^2 = \Sigma_{i=1,n}SD[X_i]^2$

For positively correlated variables having uncertain or variable p[5], an assumed value of p=½ could be argued as representative because it is the maximum entropy estimate of uncertain/variable p values equally distributed between 0 and 1.[6] For a condition where the correlation is unknown and could range with equal probability between –1 and +1, p=0 would be the maximum entropy estimate. A relationship similar to Eq 15 can be easily developed from Eq 8 for the special case that all $p_{i,j}$ are negative and equal to p.

Products of Uncertain Variables

Products of uncertain variables are also important, where the product is defined as follows:

$$\begin{aligned} X &= J_{i=1,n}\, X_i \\ &= X_1 C X_2 C \ldots C X_n \end{aligned} \quad (16)$$

[5] And in the absence of information to the contrary.
[6] The maximum entropy estimate being the midpoint of the range.

Unlike linear functions, the functional relationships for products generally depend on whether or not the X_i variables are independent (uncorrelated) or dependent (correlated). The particular case that the X_i are independent is discussed first.

Products of Independent Variables

For the particular case that the uncertain variables are independent (i.e., not correlated), the following expressions for the probabilistic parameters are exact (Benjamin and Cornell 1970):

$$E[X] = E[\prod_{i=1,n} X_i] = \prod_{i=1,n} E[X_i] \quad (17)$$

$$CV[X] = V[X]^{1/2} / E[X] = \{ \prod_{i=1,n} (CV[X_i]^2+1) - 1\}^{1/2} \quad (18)$$

Eqs 17 and 18 are only mathematically exact if all the correlation coefficient between the X_i are zero. For this case, E[X] and CV[X] do not depend on the PDFs of the X_i.

Products of Dependent Variables

Unlike the mathematically exact probabilistic parameters of linear functions of correlated uncertain variables, which do not depend on the PDFs of the variables, products, and quotients, correlated uncertain variables do generally depend on the PDFs of the individual variables.[7] Mathematically exact probabilistic parameters for products and quotients of uncertain variables, as generally used in the analysis, were based on lognormal distributions, as discussed in the Section A.2.2.

Although mathematically exact estimates for products of correlated variables require knowledge of the PDFs, general mathematical approximations for products or more complex correlated or uncorrelated functions, do not. Various approximation techniques, such as Taylor Series expansions, can be used to approximate the expected values, variances (and therefore coefficients of variation), and covariances of any (generally continuous) functions, with or without correlation between the variables. Taylor Series expansion techniques are presented by Benjamin and Cornell (1970), Meyer (1975), Hahn and Shaprio (1967), and Kaplan (1984). Various other techniques are also available (e.g., Harr 1987).

[7] Approximations based on Taylor Series expansions do not require PDFs.

A.2.2 USE OF LOGNORMAL PROBABILITY DENSITY FUNCTIONS

Uncertainties in the physical variables of interest in the analysis are generally represented by lognormal PDFs. Lognormal PDFs are considered reasonable for empirical, practical, and theoretical reasons. Empirically, the lognormal distribution is a pattern commonly found in the natural world. More specifically:

- Measured discharges and chemical concentrations and calculated loading rates observed in the Coeur d'Alene Basin generally appear lognormally distributed, as discussed in the RI/FS (USEPA 2000a and 2000b).

- Measured concentration of trace chemicals in various natural media[8] (e.g., Ott 1995; Gilbert 1987; USEPA 1992a, 1992b, 1989; Jornel and Huijbregts 1978; Krumbein and Graybill 1965) and various physical parameters commonly appear lognormally distributed (e.g., hydraulic conductivity, particle-size distributions, areas or volumes of alluvium).

Practical reasons for using lognormal distributions include the following:

- Minimum statistical assumptions are required.[9]

- The mathematics of products, quotients, and power series can be handled exactly without loss of mathematical rigor or accuracy.

- Negative values of physical variables, which are impossible, are prevented.

- Unbounded positive values are allowed (which is generally conservative because it tends to overestimate true values).

- Mixed lognormal PDFs can be treated as piece-wise lognormal PDFs.

- Any PDF can be conservatively approximated using a lognormal PDF that "envelopes" the PDF over the range of interest.

[8] Including surface water, groundwater, soil, sediments, rock, and ore bodies.
[9] A lognormal PDF is believed to be a *maximum entropy* PDF for the log of variables where only the expected value and coefficient of variation of the distribution is known or estimated. Maximum entropy estimates give the "least prejudiced, or least biased, assignment of probabilities" (Harr 1987).

The theoretical explanation as to why metal concentrations and loadings, in particular, should follow lognormal distributions comes from the physics and mathematics of probability and random processes, including the theory of successive random dilutions, the law of proportional effect, and the central limit theorem (see Ott 1995). Lognormal PDFs theoretically follow from the central limit theorem for phenomena arising from multiplicative mechanisms acting on a number of factors (Benjamin and Cornell 1970).

For practical purposes, sums of correlated variables following lognormal distributions are also lognormal, based on quasi-theoretical arguments and empirical evidence. First, lognormal PDFs do not violate the central limit theorem for phenomena arising from the sum of a large number of correlated variables that are lognormally distributed. One important empirical demonstration of this effect comes from mining, where (e.g., Jornel and Huijbregts 1978):

> Metal concentrations in ore body samples continue to follow lognormal distributions as the spatial scale of the samples increase from drill-core sized to mining-block sized, which represent the volumetric sum of hundreds or thousands of core-sized samples[10]

The most relevant and compelling empirical demonstration comes directly from basin streams, where:

> Metal concentrations and loadings in streams continue to follow lognormal distributions as the loadings from successive streams merge and add at downstream locations

This empirical demonstration for metal concentrations in streams is consistent with relevant theory, based on successive random dilutions, the law of proportional effect, and the central limit theorem.

In addition, although, theoretically, the sum of independent lognormal distributions is not lognormal,[11] it can be demonstrated by simulation that the sum closely approximates a lognormal PDF. Therefore, the sum of independent lognormal distributions can also be *approximated* as lognormal. This result is consistent with observations that loadings from multiple sources are lognormal. It is concluded for the purpose of the analysis that the sum of correlated or uncorrelated variables that are lognormally distributed can be adequately approximated as lognormally distributed.

---

[10] This phenomenon goes by the name "conservation of lognormality" in mining geostatistics.

[11] By the central limit theorem, as the number of variables in the sum increases, the sum should approach a normal distribution. Note that the difference between a normal PDF and lognormal PDF is proportional to the coefficient of variation. At small values of the coefficient of variation, the difference is minimal.

A.2.2.1 Lognormal Probability Functions and Parameters

The following relationships assume that the uncertainty in the true value of the variable of interest, X, is lognormally distributed. Symbol "x" represents any given specific estimate of (true) X. The relationships are mathematically exact and adapted from Benjamin and Cornell (1970).

The probability density function of X for any given estimate of X, x, is symbolized as $PDF_X[x]$ and calculated as follows:

$$PDF_X[x] = PDF_U\{u\} / x \ln\{CV[X]^2 + 1\}^{1/2} \quad (1)$$

where:

$PDF_U\{u\}$	=	Standard normal probability density function for X evaluated at u
u	=	Standard normal variate for a lognormal variable
	=	$\ln\{x\ (CV[X]^2+1)^{1/2}/E[X]\}\ln(CV[X]^2+1)^{-1/2}$
	=	$\ln\{x\ /M[X]\}\ln(CV[X]^2+1)^{-1/2}$
E[X]	=	Expected value of X
CV[X]	=	Coefficient of variation of X
M[X]	=	Median of X
	=	$E[X] / (CV[X]^2+1)^{1/2}$
x	=	Any estimate of X

The probability that X lies in a region bounded between a lower limit x_a and an upper limit x_b is the integral of $PDF_X[x]dx$ from x_a to x_b. The cumulative density function of X, $CDF_X[x]$, is the integral of $PDF_X[x]dx$ from x_a=0 to x_b=x and measures the probability that X is less than or equal to the estimate x; that is:

$$CDF_X[x] = P[X \leq x] \quad (2)$$
$$= CDF_U\{u\}$$

where $CDF_U\{u\}$ is the standard normal cumulative distribution function of X evaluated at u and all other parameters are as defined for $PDF_X[x]$. Values of $CDF_U\{u\}$can be obtained from standard statistical tables (e.g., Benjamin and Cornell 1970) or standard spread sheet functions

(e.g., function NORMSDIST(u) in MS Excel 97). Note that for the special case of x=E[X], $P[X \leq E[X]] = CDF_U\{u = (\ln\{CV[X]^2+1\})^{1/2}\}$.

The complement of $CDF_X[x]$ is the probability that X *exceeds* x; that is:

$$1 - CDF_X[x] = P[X>x] = 1 - CDF_U\{u\} \quad (3)$$

The products and quotients of lognormally distributed variables make use of the expected value and variance of the natural log of X, E[lnX] and V[lnX]. These parameters are related to the expected value and coefficient of variation of X, E[X] and CV[X], by the following relationships:

$$E[\ln X] = \ln\{E[X]/(CV[X]^2+1)^{1/2}\} \quad (4)$$

$$V[\ln X] = \ln\{CV[X]^2+1\} \quad (5)$$

The following section addresses techniques to estimate E[X] and CV[X] from available data.

A.2.2.2 Lognormal Parameter Estimates From Statistical Data Analysis

The two probabilistic parameters of a lognormal distribution, E[X] and CV[X], can be estimated from available measurement data using a variety of statistical methods. Two common methods are presented here.

Statistical Moments

The first method for estimating E[X] and CV[X] is based on statistical "moments" of the data. The method of moments can be based on ln-transformed or untransformed data. For untransformed data E[X] and CV[X] are based on the arithmetic mean (m_X) and standard deviation (s_X) or variance (s_X^2) of the data (x_i, i=1 to n, where n is the total number of data points); that is:

$$E[X] = (\Sigma_{i=1,n}\, x_i) / n = m_X \quad (1)$$

$$CV[X] = s_X / m_X \quad (2)$$

$$= \{s_X^2\}^{1/2} / m_X$$
$$= \{[\Sigma_{i=1,n} (x_i - m_X)^2] / n\text{-}1\}^{1/2} / m_X$$

Statistical Error Associated with Expected Value Estimates

Statistical (random) error associated with E[X] estimated from n independent measurement data can be approximated as follows:

$$CV[E[X]] = (s_X / m_X) / n^{½} \quad (3)$$

Note that (s_X / m_X) is the same as the estimate of CV[X] computed by Eq 2. Eq 3 assumes that estimate E[X] by Eq 1 is normally distributed. The validity of this assumption is based on the central limit theorem and increases with n, the number of samples used in the estimate. As Natrella (1966) states, ". . . the sampling distribution of E[X][12] is "normal" to a good approximation for samples of four or more measurements from almost every population distribution likely to be met in practice" (Natrella 1966, pp. 2-9 and 2-10). This approximation is consistent with the common approach for estimating confidence intervals or limits around the true mean, or E[X], computed from measurement data. Specifically, the 1-" two-sided confidence interval for the expected value, E[X], is computed as follows:

$$m_X - (s_X/n^{½})t_{"/2,n-1} \leq E[X] \leq m_X + (s_X/n^{½})t_{"/2,n-1} \quad (4)$$

where $t_{"/2,n-1}$ is the t-statistic at the 1-"/2 confidence level with n-1 degrees of freedom (Natrella 1966).[13] Based on Eq 4, the 1-"% one-sided upper confidence level (UCL) on the expected value would be $m_X+(s_X/n^{½})t_{",n-1}$; similarly, the 1-"% one-sided lower confidence level (LCL) on the expected value would be $m_X- (s_X/n^{½})t_{",n-1}$. Statistical approaches to parameter estimation and comparison are found in standard statistical handbooks and texts (e.g., Natrella 1966, Benjamin and Cornell 1970, Meyer 1975).

[12] Natrella uses X with a bar over it for E[X].
[13] It is noted that Gilbert (1987, p.170) states that if one is confident that the underlying distribution is lognormal, the method of Land is preferred. However, the method of Land is considered suspect for practical cases because of its extreme predictions. In most practical cases, physical realizations in the extreme tail of a theoretical lognormal distribution having large coefficients of variation are not reasonably possible. These suspect, extreme values can control and high-bias the expected value.

For X lognormally distributed, lnX follows a normal distribution, as would ln-transformed data. E[X] and CV[X] can be estimated from the mean (m_{lnX}) and variance (s_{lnX}^2) of the ln-transformed data using Eqs 1 and 2 (calculated on the ln-transformed data) and the following relationships (Gilbert 1987, Benjamin and Cornell 1970):

$$E[X] = \exp\{m_{lnX} + s_{lnX}^2/2\} \quad (5)$$

$$CV[X] = V[X]^{1/2} / E[X]$$
$$= \{\exp(2m_{lnX} + s_{lnX}^2)[\exp(s_{lnX}^2) - 1]\}^{1/2} / E[X] \quad (6)$$

Also, the geometric mean or median of X, M[X], is estimated as $\exp\{m_{lnX}\}$. Similar to computing bounds around E[X], the 1-" two-sided confidence interval for the geometric mean or median of X, M[X], can be computed as follows (Gilbert 1987, p.173):

$$\exp\{m_{lnX} - (s_{XnX}/n^{½})t_{"/2,n-1}\} \leq M[X] \leq \exp\{m_{lnX} + (s_{lnX}/n^{½})t_{"/2,n-1}\} \quad (7)$$

Based on Eq 7, the 1-"% one-sided UCL on the geometric mean or median would be $\exp\{m_{lnX} + (s_{lnX}/n^{½})t_{",n-1}\}$; the similar LCL would be $\exp\{m_{lnX} - (s_{lnX}/n^{½})t_{",n-1}\}$.

Regression Analysis

The second method for estimating E[X] and CV[X] is based on linear least-squares regression analysis (Ecology 1993). The normal standard variate (u) is regressed against the natural log of the data using the linear model u=b+m*ln{x}, where m and b are the slope and intercept, respectively, of the regression line. Using this method, the data are ranked from lowest to highest (i=1 to n) and assigned a "plotting point," which is an estimate of the cumulative frequency or probability of each data (x_i or $\ln\{x_i\}$). The plotting point, PP, (for a normal distribution) for each x_i is as follows, where i is the rank of x_i:

$$PP_i = (i - 3/8) / (n + 1/4) \quad (8)$$

The value of u for each x_i is then calculated as the inverse standard normal cumulative distribution, CDF^{-1}, evaluated at the plotting point; that is:

$$u_i = CDF^{-1}\{(i - 3/8) / (n + 1/4)\} \quad (9)$$
$$= CDF^{-1}\{PP_i\}$$

The u_i are regressed against $\ln\{x_i\}$to calculate the least squares ("best fit") regression line, $u=b+m*\ln\{x\}$. The square of the correlation coefficient, r^2, measures the goodness of fit to a straight line, which is indicative of lognormally distributed data (e.g., as a test for lognormality, Ecology (1993) assumes the data are lognormally distributed for $r^2 \geq 0.9$).

E[X] and CV[X] are estimated from the slope and intercept of the regression line, $u=b+m*\ln\{x\}$, using the following relationships for a lognormal variate (Ecology 1993, Gilbert 1987):

$$E[X] = \exp[-b/m + ½m^{-2}] \quad (10)$$

$$CV[X] = \{\exp[m^{-2}] - 1\}^{1/2} \quad (11)$$

The regression method was generally used in the RI to analyze the available historic data to estimate the expected value and coefficient of variation of discharge, metal concentration, and metal loading at selected stream monitoring locations. Further discussion is presented in Appendix B Section B.3.1.

Note that E[X] estimated by regression depends on both the slope and intercept of the regression line. Because of this, there appears to be no direct way to estimate confidence intervals for statistical uncertainty in E[X] estimated by regression. However, error bounds can be approximated using results from the method of statistical moments; that is, following Eq 4:

$$\exp[-b/m + 1/2m^2] - (s_X/n^{½})t_{"/2,n-1} \leq E[X] \leq \exp[-b/m + 1/2m^2] + (s_X/n^{½})t_{"/2,n-1} \quad (12)$$

Based on Eq 12, the 1-"% one-sided UCL on E[X] would be $\exp[-b/m + 1/2m^2] + (s_X/n^{½})t_{",n-1}$. The similar LCL would be $\exp[-b/m + 1/2m^2] - (s_X/n^{½})t_{",n-1}$.

The following two sections derive relationships for the expected value and coefficient of variation of variables that are functions of lognormal variables. Section A.2.3.2 covers products and Section A.2.3.3 covers quotients.

A.2.2.3 Probabilistic Parameters for Products of Lognormal Variables

The product of two lognormally distributed random variables is also lognormally distributed (Benjamin and Cornell 1970). The expressions for the coefficient of variation of the product of two lognormally distributed random variables X=YCZ is derived as follows:

$$X = YCZ \tag{1}$$

$$\ln X = \ln Y + \ln Z \tag{2}$$

$$\begin{aligned} V[\ln X] &= V[\ln Y + \ln Z] \\ &= V[\ln Y] + V[\ln Z] + 2Cov[\ln Y, \ln Z] \\ &= V[\ln Y] + V[\ln Z] + 2\, p_{\ln Y,\ln Z}(V[\ln Y]V[\ln Z])^{1/2} \end{aligned} \tag{3}$$

Note that Eq 3 follows from Eq 7 in Section A2.1.1. Using "A" as a dummy variable for X, Y, or Z, and substituting $V[\ln A]=\ln\{CV[A]^2+1\}$ for all variables and letting $p=p_{\ln Y,\ln Z}$ yields:

$$\begin{aligned} \ln(CV[X]^2+1) &= \ln(CV[Y]^2+1)+\ln(CV[Z]^2+1) + 2p(V[\ln Y]V[\ln Z])^{1/2} \\ &= \ln(CV[Y]^2+1)+\ln(CV[Z]^2+1) + \\ &\quad + 2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2} \end{aligned} \tag{4}$$

Taking anti-natural logs and simplifying:

$$CV[X]^2+1 = (CV[Y]^2+1)(CV[Z]^2+1)\exp\{2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\} \tag{5}$$

$$\begin{aligned} CV[X] &= \{(CV[Y]^2+1)(CV[Z]^2+1)\exp\{2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\}-1\}^{1/2} \\ &= \{(CV[Y]^2+1)(CV[Z]^2+1)\mathbf{S}^2 - 1\}^{1/2} \end{aligned} \tag{6}$$

where:

$$\begin{aligned} \mathbf{S} &= \exp\{2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\}^{1/2} \\ &= \exp\{p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\} \end{aligned}$$

p = Correlation coefficient between the natural log of Y and the natural log of Z, $p_{\ln Y,\ln Z}$

The expression for expected value of the product of two lognormally distributed random variables X=Y℄Z is derived as follows:

$$E[\ln X] = E[\ln Y + \ln Z] \qquad (7)$$
$$= E[\ln Y] + E[\ln Z]$$

Substituting $E[\ln A]=\ln\{E[A]/(CV[A]^2+1)^{1/2}\}$ for each of the variables (where A is a dummy variable for X, Y, or Z), taking antilogs, and algebraically simplifying leads to the following expressions:

$$\ln\{E[X]/(CV[X]^2+1)^{1/2}\} = \ln\{E[Y]/(CV[Y]^2+1)^{1/2}\} + \ln\{E[Z]/(CV[Z]^2+1)^{1/2}\} \qquad (8)$$

$$E[X]/(CV[X]^2+1)^{1/2} = \{E[Y]/(CV[Y]^2+1)^{1/2}\}\{E[Z]/(CV[Z]^2+1)^{1/2}\} \qquad (9)$$

$$E[X] = E[Y]E[Z]/\{(CV[X]^2+1)^{1/2}(CV[Z]^2+1)^{1/2}\}(CV[X]^2+1)^{1/2} \qquad (10)$$

Substituting for CV[X] from Eq 6 and algebraically simplifying yields the following, where $\mathbf{S}$ is as before for CV[X]; that is:

$$E[X] = E[Y]E[Z]\mathbf{S} \qquad (11)$$

To summarize, the product X=Y℄Z of two lognormally distributed random variables is also lognormally distributed with expected value and coefficient of variation given by the following expressions:

$$E[X] = E[Y]E[Z]\mathbf{S} \qquad (12)$$

$$CV[X] = \mathbf{\{}(CV[Y]^2+1)(CV[Z]^2+1)\mathbf{S}^2 - 1\mathbf{\}}^{1/2} \qquad (13)$$

where:

$\mathbf{S}$ = $\exp\mathbf{\{}2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\mathbf{\}}^{1/2}$
= $\exp\mathbf{\{}p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\mathbf{\}}$

p = Correlation coefficient between the natural log of Y and the natural log of Z, $p_{\ln Y,\ln Z}$

For X=YCZ, both E[X] and CV[X] increase with the correlation coefficient, from a minimum if Y and Z are perfectly negatively correlated, such that $p_{lnY,lnZ}$=–1, to a maximum if Y and Z are perfectly positively correlated, such that $p_{lnY,lnZ}$=1. Note that if $p_{lnY,lnZ}$=0, then **S**=1 in Eqs 5 and 6.

Generalizing the products of lognormal variables for X = $\mathrm{J}_{i=1,n} X_i$, results in the following:

$$E[X] = \{ \mathrm{J}_{i=1,n} E[X_i]\} \mathbf{S} \quad (14)$$

$$CV[X] = \{\{ \mathrm{J}_{i=1,n}(CV[X_i]^2+1)\} \mathbf{S}^2 - 1\}^{1/2} \quad (15)$$

where:

$$\mathbf{S} = \exp\{2\Sigma_{i=1,n}\Sigma_{j=i+1}p_{i,j}\{\ln(CV[X_i]^2+1)\ln(CV[X_j]^2+1)\}^{1/2}\}^{1/2}$$
$$= \exp\{\Sigma_{i=1,n}\Sigma_{j=i+1}p_{i,j}\{\ln(CV[X_i]^2+1)\ln(CV[X_j]^2+1)\}^{1/2}\}$$

p = Correlation coefficient between the natural log of X_i and the natural log of X_j, $p_{lnXi,lnXj}$

If $p_{lnXi,lnXj}$=0 for all i and j, then **S**=1 and Eqs 14 and 15 simplify to Eqs 17 and 18 in Section A.2.1.3, which are mathematically exact for uncorrelated variables having any PDF, not just lognormal.

A.2.2.4 Probabilistic Parameters for Quotients of Lognormal Variables

The quotient of two lognormally distributed random variables is also lognormally distributed (Benjamin and Cornell 1970). The expressions for the coefficient of variation of the quotient of two lognormally distributed random variables X=Y/Z follows the same general derivation as for the product of two random variables, with the following differences:

$$X = Y/Z \quad (1)$$

$$\ln X = \ln Y - \ln Z \quad (2)$$

$$E[\ln X] = E[\ln Y - \ln Z] \quad (3)$$
$$= E[\ln Y] - E[\ln Z]$$

$$V[\ln X] = V[\ln Y - \ln Z] \quad (4)$$
$$= V[\ln Y] + V[\ln Z] - 2Cov[\ln Y,\ln Z]$$

$$= \quad V[\ln Y] + V[\ln Z] - 2p_{\ln Y,\ln Z}(V[\ln Y]V[\ln Z])^{1/2}$$

Compared to X=YCZ, for X=Y/Z, lnZ is *subtracted* from lnY; Eq 4 then follows from Eq 7 in Section A.2.1.1. After substitutions and algebraic simplifications similar to those used for X=YCZ in Section A.2.3.1, the following results are obtained for X=Y/Z:

$$E[X] = E[Y](1+ CV[Z]^2)\mathbf{S} / E[Z] \quad (5)$$

$$CV[X] = \{(CV[Y]^2+1)(CV[Z]^2+1)\mathbf{S}^2 - 1\}^{1/2} \quad (6)$$

where:

$$\mathbf{S} = \exp\{-2p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\}^{1/2}$$
$$= \exp\{-p\{\ln(CV[Y]^2+1)\ln(CV[Z]^2+1)\}^{1/2}\}$$

p = Correlation coefficient between the natural log of Y and the natural log of Z, $p_{\ln Y,\ln Z}$

Note the differences compared to the product of two uncertain variables. First, the expected value of the quotient depends on the coefficient of variation of the denominator. Second, the correlation coefficient p in **S** is multiplied by a –1 (i.e., –p). In contrast to the product of two variables, for X=Y/Z, both E[X] and CV[X] *decrease* with the correlation coefficient, from a maximum if Y and Z are perfectly negatively correlated, such that $p_{\ln Y,\ln Z} = -1$, to a minimum if Y and Z are perfectly positively correlated, such that $p_{\ln Y,\ln Z} = 1$. Note that if $p_{\ln Y,\ln Z}=0$, then **S**=1 in Eqs 5 and 6.

A.2.2.5 Probabilistic Parameters for Power Series of Uncertain Variables

Generalizing the products and quotients of lognormal variables for a power series, where the exponents a_i are constants that can be positive (products, $a_i>0$) or negative (quotients, $a_i<0$), results in the following:

$$X = \mathbf{J}_{i=1,n} X_i^{ai}, \quad (1)$$

$$E[X] = \{\mathbf{J}_{i=1,n} E[X_i]^{ai}(1+ CV[X_i']^2)^{ai*ai}\}\mathbf{S} \quad (2)$$

$$CV[X] = \{\{\mathbf{J}_{i=1,n}(CV[X_i]^2+1)^{ai*ai}\}\mathbf{S}^2 - 1\}^{1/2} \quad (3)$$

where:

X_i'	=	X_i with negative a_i; for positive a_i, $X_i' = 0$
$a_i * a_i$	=	a_i^2
S	=	$\exp\{2\Sigma_{i=1,n}\Sigma_{j=i+1}p_{i,j}a_ia_j\{\ln(CV[X_i]^2+1)\ln(CV[X_j]^2+1)\}^{1/2}\}^{1/2}$
	=	$\exp\{\Sigma_{i=1,n}\Sigma_{j=i+1}p_{i,j}a_ia_j\{\ln(CV[X_i]^2+1)\ln(CV[X_j]^2+1)\}^{1/2}\}$
p	=	Correlation coefficient between the natural log of X_i and the natural log of X_j, $p_{lnXi,lnXj}$

Eqs 2 and 3 are not generally mathematically exact for non-lognormal variables (unless all a_i=1), although the approximation would generally be of high-order, as indicated by comparison with the following Taylor series approximation for uncertain variables following any PDF, not restricted to lognormal (or any) distribution.

Unrestricted PDFs

A second-order Taylor series approximation for a power series composed of uncertain variables each having *any* PDF is as follows (Meyer 1975, Hahn and Shaprio 1976, Benjamin and Cornell 1970, and Kaplan 1984):

$$E[X] = E[\prod_{i=1,n} X_i^{ai}] \tag{4}$$

$$\simeq \prod_{i=1,n} E[X_i]^{ai} + \tfrac{1}{2} \sum_{i=1,n}\sum_{j=1,n} (\partial^2 X/\partial X_i \partial X_j|_m)Cov[X_i,X_j]$$

$$\simeq \prod_{i=1,n} E[X_i]^{ai} \Big(1 + \tfrac{1}{2} \sum_{i=1,n} a_i(a_i-1)CV[X_i]^2 + $$

$$+ \sum_{i=1,n}\sum_{j=i+1,n} p_{Xi,Xj}\, a_ia_j\, CV[X_i]\, CV[X_j]\Big)$$

$$V[X] = V[\prod_{i=1,n} X_i^{ai}] \tag{5}$$

$$\simeq \sum_{i=1,n}\sum_{j=1,n} \Big\{(\partial X/\partial X_i|_m\, \partial X/\partial X_j|_m)Cov[X_i,X_j] + (\partial^2 X/\partial X_i \partial X_j|_m)V[X_i]V[X_j]\Big\}$$

$$\simeq (\prod_{i=1,n}E[X_i]^{ai})^2\Big(\sum_{i=1,n} a_i^2\, CV[X_i]^2 + 2\sum_{i=1,n}\sum_{j=i+1,n} p_{Xi,Xj}\, a_ia_j\, CV[X_i]CV[X_j]\Big) +$$

$$+ \Big\{(\prod_{i=1,n} E[X_i]^{ai})\sum_{i=1,n}\sum_{j=i+1,n} a_ia_j\, E[X_i]E[X_j]CV[X_i]CV[X_j]\Big\}$$

$$CV[X] = V[X]^{1/2} / E[X] \quad (6)$$

$$\simeq \left\{\prod_{i=1,n}(CV[X_i]^2 + 1) - 1 + 2\sum_{i=1,n}\sum_{j=i+1,n} \rho_{Xi,Xj}\, a_i a_j\, CV[X_i]CV[X_j]\right\}^{1/2}$$

$$\simeq \left\{(CV[X_1]^2+1)(CV[X_2]^2+1)...(CV[X_n]^2+1) - 1 + \right.$$

$$\left. + 2\sum_{i=1,n}\sum_{j=i+1,n} \rho_{Xi,Xj}\, a_i a_j\, CV[X_i]CV[X_j]\right\}^{1/2}$$

A.2.2.6 Probabilistic Parameters for a General Function of a Lognormal Variable

In general, the expected value and coefficient of variation of a function X=f(Y) where Y is lognormally distributed can be estimate from the fundamental definitions of expected values and variances (Benjamin and Cornell 1970); that is:

$$E[X] = {}_{0\text{ to }4}\, f(y)PDF[Y]\, dy \quad (1)$$

$$CV[Y] = V[Y]^{1/2} / E[Y] \quad (2)$$

$$= \{ {}_{0\text{ to }4}\, f(y)^2\, PDF[Y]\, dy - E[X]^2\}^{1/2} / E[X]$$

Typically, Eqs 1 and 2 have to be solved numerically.

A.3.0 PROBABILISTIC ESTIMATES

This section provides the mathematical relationships required to make probabilistic estimates. The probability that X exceeds some given value (e.g., "G") is covered in Section A.3.1. The non-exceedance value of X, the value of X that is not exceeded with a given probability (e.g., 95 percent), is covered in Section A.3.2. The value of X required to meet a given value (e.g., G) with some specified probability is covered in Section A.3.3. As discussed in Section A.2, these probabilistic estimates assume that the uncertainty in X follows a lognormal PDF.

A.3.1 PROBABILITY TRUE VALUE OF X DOES NOT EXCEED GOAL G

The probability that (the true but uncertain value of) X does not exceed a given target value or goal G[14] is estimated as follows from the probabilistic parameters of X (Benjamin and Cornell 1970); that is:

$$P[X \leq G] = CDF_X\{u\} \qquad (1)$$

where:

$CDF_X\{u\}$	=	Standard normal cumulative distribution function of X evaluated at u
u	=	Standard normal variate for a lognormal variable
	=	$\ln\{G(CV[X]^2+1)^{1/2}/E[X]\}\ln(CV[X]^2+1)^{-1/2}$
	=	$\ln\{G/M[X]\}\ln(CV[X]^2+1)^{-1/2}$
E[X]	=	Expected value of X
CV[X]	=	Coefficient of variation of X
M[X]	=	Median of X
G	=	Target value or goal for X
X	=	True (exact but uncertain) value of X

Eq 1 assumes that the uncertainty in X follows a lognormal PDF (and thus CDF). Values of $CDF_X\{u\}$ can be obtained from statistical tables for standard normal variates (e.g., Benjamin and Cornell 1970) or from a standard spreadsheet function (e.g., MS Excel 97). The probability that X *exceeds* G is the complement of Eq 1; that is:

[14] E.g., X may be mass loading and G may be a TMDL loading capacity.

$$P[X > G] = 1 - CDF_X\{u\} = 1 - P[X \leq G]$$

Eq 1 can provide a quantitative answer to the important, practical question: *What is the probability of meeting cleanup goal G if X, the quantity used to measure cleanup, has a given estimated value and degree of uncertainty?* The answer is illustrated in Figure A-1, which quantitatively demonstrates that the practical attainment of a cleanup goal depends on *both* the magnitude *and* uncertainty in the quantity used to measure cleanup attainment.

Specifically, using Eq 1, Figure A-1 quantifies how the probability of achieving a cleanup goal, G, varies with the expected value, E[X], and degree of uncertainty, measured by the coefficient of variation, CV[X], of variable X, used to measure cleanup attainment. For selected value of CV[X], Figure A-1 graphs the relationship between (1) E[X]/G, the ratio of the expected value of X and the goal G, and (2) P[X<G], the probability that the true (but uncertain) value of X is below goal G. The ratio E[X]/G is the number of times the expected value of X exceeds goal G (e.g., for E[X]/G = 10, E[X] is 10 times G; for E[X]/G = 1, E[X] is equal to G; and so on).

Figure A-1 also illustrates the classic "two-edged sword" of uncertainty. For relatively low uncertainty (e.g., CV[X]=0.10), the probability that the cleanup goal has been met is negligible until the expected value of X is nearly equal to the goal. On the other hand, for relatively high uncertainty (e.g., CV[X]=1.0), there is at least a small probability that the goal is met even if the expected value of X is many times the goal. However, to achieve a relatively high (e.g., 90 percent) probability that the goal has been met, the expected value of X must be *below* G, with the magnitude of the difference increasing with the degree of uncertainty in X.

A.3.2 NON-EXCEEDANCE ESTIMATES OF X, X_{PN}

Estimates having a suitably high probability of being *underestimates* can also be useful in assessing current conditions and potential future conditions. Estimates having a specified probability of being underestimates are known as "non-exceedance estimates."

Non-exceedance estimates of X have a given probability, termed the "non-exceedance probability," *of not being exceeded* by the true but uncertain value of X. For example, a 90 percent non-exceedance estimate of X has a 90 percent probability of being an underestimate—that is, of not being exceeded by the true value of X.

Non-exceedance estimates having a relatively high probability of non-exceedance (e.g., $\geq$90 percent), are *"reliable"* estimates—reliable in that there is a low probability (e.g, $\leq$10 percent) that the true (but uncertain) value of X exceeds the estimate. It is in this sense that

a probabilistic analysis can be used to make reliable estimates—by using a suitably high probability of non-exceedance.

Non-exceedance probabilities are denoted "Pn," with the non-exceedance estimate denoted as x_{Pn}. For given non-exceedance probability Pn, x_{Pn} is obtained using the lognormal standardized normal variate "u" corresponding to Pn: u_{Pn}. As an example, for a 90 percent probability of non-exceedance, Pn=90 percent , and u_{Pn}=1.28. The non-exceedance probability Pn is the probability that x_{Pn} is not an underestimate of the true value of X. For a lognormally distributed X, x_{Pn}, is given by the following, derived from Eq 1; that is:

$$\begin{aligned} x_{Pn} &= E[X]\exp\{u_{Pn}\ln(CV[X]^2+1)^{1/2}\} / (CV[X]^2+1)^{1/2} \\ &= M[X]\exp\{u_{Pn}\ln(CV[X]^2+1)^{1/2}\} \\ &= \text{Estimate of x having } P[X \leq x_{Pn}]=Pn \end{aligned} \quad (2)$$

For illustrative convenience, a non-exceedance ratio, $x_{Pn}/E[X]$, can be defined from Eq 2. The non-exceedance ratio is the number of times that the non-exceedance estimate, x_{Pn}, exceeds the expected value, E[X]. This ratio is illustrated and quantified in Figure A-2 for selected levels of uncertainty in X, measured by CV[X].

Figure A-2 shows that the non-exceedance ratio is less than 1.0 for low values of the non-exceedance probability, Pn, and grater than 1.0 for high values of Pn; and that the range of ratios increases with the uncertainty in X, measured by CV[X]. For high values of Pn (e.g., $\geq$90 percent) and uncertainty (e.g., CV[X]>0.50), Figure A-2 shows that the non-exceedance estimate, x_{Pn}, is approximately two to three times the expected value, E[X].

In practical terms, Figure A-2 demonstrates the potential for underestimating an uncertain quantity, whether predicted or measured. If there is significant uncertainty, a reliable estimate would exceed the expected value (or best estimate) by two to three times.

Exceedance values can also be estimated. An exceedance estimate of X has a given probability, termed the "exceedance probability" and denoted "Pe," of *being* exceeded by the true value of X. An exceedance estimate of X, denoted "x_{Pe}," has a probability equal to Pe of being exceeded by X. An exceedance estimate at a given probability level Pe is the non-exceedance estimate at the complementary probability level—i.e., Pe=1-Pn. For example, a 90 percent *exceedance* estimate is also a 10 percent *non*-exceedance estimate; and vice versa. That is:

$$x_{Pe} = E[X]\exp\{-u_{Pn}\ln(CV[X]^2+1)^{1/2}\} / (CV[X]^2+1)^{1/2}$$

= Estimate having $P[X > x_{Pe}] = Pe = 1 - Pn$

= $x_{Pn=1-Pe}$

Note that for normal standard variates, u, at given probability level: $u_{Pe} = -u_{Pn}$, where, again, $Pe = 1 - Pn$.

A.3.3 ESTIMATE OF X REQUIRED TO MEET GOAL G WITH SPECIFIED PROBABILITY

The quantity estimated is the expected value of X required to meet goal G with specified probability, Pr, and given uncertainty in X, measured by CV[X]. Symbolized $E[X_G]$, this quantity is equivalent to the E[X] required for P[X<G]=Pr. From Eq 1, $E[X_G]$ is estimated as follows, where u_{Pr} is the standard normal variate corresponding to specified probability Pr; that is:

$$E[X_G] = G\,(CV[X]^2+1)^{1/2} / \exp\{u_{Pr} \ln(CV[X]^2+1)^{1/2}\} \qquad (3)$$

Eq 3 is illustrated by Figure A-3, which (not surprisingly) looks like Figure A-1. Figure A-3 graphs the ratio of $E[X_G]/G$ against Pr, the specified probability, over the range of 0.50 to 0.99 for various values of CV[X], which measures the uncertainty in X. Consistent with the implications of Figure A-1, Figure A-3 indicates that $E[X_G]$ is less than G at higher values of Pr, with the magnitude of the difference proportional to CV[X]. That is, achieving a high probability that the true value of X is less than G requires that the expected value of X be *less* than G, with the magnitude of the difference increasing with the uncertainty in X.

A.4.0 BAYESIAN ESTIMATES OF EXPECTED VALUES AND COEFFICIENTS OF VARIATION

Expected values and coefficients of variation of variables of interest are estimated from available measurement data to the extent data are available. In particular, estimates for pre-remediation metal loadings are based on analysis of available historic stream monitoring data, as discussed and presented in the RI/FS (USEPA 2000a and 2000b).

More generally, and particularly for post-remediation conditions, a Bayesian approach is used to estimate probabilistic parameters using weight-of-evidence arguments and professional judgment. A Bayesian approach allows probabilistic parameters (or any variable) to be estimated using any relevant and appropriate means, including:

- Available measurement and other observational data
- Exact or approximate relationships between X and other variables (i.e., where X is a deterministic or probabilistic function of other variables)
- Information from similar or related situations
- Professional judgment, including expert opinion

Professional judgments and interpretations are documented and quantified, as scientifically and practically appropriate.

A Bayesian approach is fundamentally a practical philosophy for estimation and decision making that allows the use of all available information, both quantitative and qualitative. Bayesian philosophy, operational theory, and practice are well supported in the mathematical, scientific, and engineering literature (e.g., Benjamin and Cornell 1970, Martz and Waller 1982, Raiffa and Schlaifer 1961, Ang and Tang 1975 and 1984, Anderson 1998). Real-world environmental applications include U.S. Navy (1997 and 2001). The scientific method itself is considered Bayesian (Martz and Waller 1982), and professional engineering practice is unavoidably Bayesian.

Where measurement data is very limited or non-existent, a Bayesian approach is used to estimate probabilistic parameters using professional interpretation and judgment. In particular, relationships based on lognormal PDFs are used to estimate E[X] and CV[X] from estimates of X at any two non-exceedance (Pn) or exceedance (Pe) probabilities.

The estimates of X are characterized at two non-exceedance or exceedance probabilities using the standard normal variate u set at $u=u_{Pn}$ for non-exceedance probabilities, Pn, and $u=u_{Pn=1-Pe}$ for exceedance probabilities, Pe; that is:

x_{u1} = Estimate of X at Probability 1
x_{u2} = Estimate of X at Probability 2

where u_1 (either u_{Pn} or $u_{Pn=1-Pe}$, as appropriate) is the standard normal variate at Probability 1, and u_2 (either u_{Pn} or $u_{Pn=1-Pe}$, as appropriate) is the standard normal variate at Probability 2. Two examples of this are:

- *There is a 95 percent probability that X will not exceed a value of 0.8 and a 95 percent probability that it will exceed a value of 0.4.* In this case, x_{u1}=0.8 at probability Pn=0.95 ($u_1=u_{Pn=.95}$) and x_{u2}=0.4 at probability Pe=0.95 ($u_2=u_{Pn=1-Pe=.05}$).
- *There is a 90 percent probability that X will not exceed a value of 1,000 and a 50 percent probability that it will exceed a value of 100.* In this case, x_{u1}=1,000 at probability Pn=0.90 ($u_1=u_{Pn=.90}$) and x_{u2}=100 at probability Pe=0.50 ($u_2=u_{Pn=1-Pe=.50}$).

The expected value and coefficient of variation of X are determined from x_{u1} and u_1 and x_{u2} and u_2 as follows:

$$CV[X] = (\exp\{[\ln\{x_{u1}/x_{u2}\}/(u_1-u_2)]^2\} - 1)^{1/2} \quad (1)$$

$$E[X] = x_{u1}\,(CV[X]^2+1)^{1/2} \,/\, \exp\{u_1\,(\ln\{CV[X]^2+1\})^{1/2}\} \quad (2)$$
$$= x_{u2}\,(CV[X]^2+1)^{1/2} \,/\, \exp\{u_2\,(\ln\{CV[X]^2+1\})^{1/2}\}$$

The median of X, M[X], can be computed from E[X] and CV[X] as follows:

$$M[X] = E[X] \,/\, (CV[X]^2+1)^{1/2}$$

Consider the useful special case that E[X] and CV[X] are estimated from "practical upper bound" (UB[X]) and "practical lower bound" (LB[X]) estimates of the true value of X. Assume that UB[X] is considered to be the 95 percent non-exceedance value and that LB[X] is

considered to be the 95 percent exceedance value (such that there is a 90 percent probability that true X lies between LB[X] and UB[X]).[15] In this case, $u_1 = -u_2 = 1.645$ and Eqs 1 and 2 become:

$$CV[X] = (\exp\{[\ln\{UB[X] / LB[X]\}/3.29]^2\} - 1)^{1/2} \quad (3)$$

$$E[X] = UB[X]\,(CV[X]^2+1)^{1/2} \,/\, \exp\{1.645\,(\ln\{CV[X]^2+1\})^{1/2}\} \quad (4)$$
$$= LB[X]\,(CV[X]^2+1)^{1/2} \,/\, \exp\{-1.645\,(\ln\{CV[X]^2+1\})^{1/2}\}$$

Figure A-4 shows the ratios UB[X]/E[X], LB[X]/E[X], and M[X]/E[X] for values of CV[X] from 0 to 2.0. Figure A-4 can be compared to Figure A-2.

An estimate of X for any other exceedance/non-exceedance probability can be determined from E[X] and CV[X] as follows, where u is the standard normal variate at the desired probability; that is:

$$x_u = E[X]\exp\{u\,(\ln\{CV[X]^2+1\})^{1/2}\} \,/\, (CV[X]^2+1)^{1/2} \quad (5)$$

Derivation of Equations 1 and 2

Derivation of Eqs 1 and 2 are as follows:

$$x_u = E[X]\exp\{u\,(\ln\{CV[X]^2+1\})^{1/2}\} \,/\, (CV[X]^2+1)^{1/2}$$

$$x_{u1} / x_{u2} = \exp\{u_1\,(\ln\{CV[X]^2+1\})^{1/2}\} \,/\, \exp\{u_2\,(\ln\{CV[X]^2+1\})^{1/2}\}$$
$$= \exp\{(u_1-u_2)\,(\ln\{CV[X]^2+1\})^{1/2}\}$$

$$\ln\{x_{u1}/x_{u2}\} = (u_1-u_2)\,(\ln\{CV[X]^2+1\})^{1/2}$$

$$\ln\{CV[X]^2+1\} = [\ln\{x_{u1}/x_{u2}\} \,/\, (u_1-u_2)]^2$$

$$CV[X]^2 = \exp\{[\ln\{x_{u1}/x_{u2}\}/(u_1-u_2)]^2\} - 1$$

[15] Similar to statistical tolerance limits (e.g., Natrella 1966).

$$CV[X] = (\exp\{[\ln\{x_{u1}/x_{u2}\}/(u_1-u_2)]^2\} - 1)^{1/2}$$

$$E[X] = x_{u1}\,(CV[X]^2+1)^{1/2} \,/\, \exp\{u_1\,(\ln\{CV[X]^2+1\})^{1/2}\}$$
$$= x_{u2}\,(CV[X]^2+1)^{1/2} \,/\, \exp\{u_2\,(\ln\{CV[X]^2+1\})^{1/2}\}$$

Martz and Waller (1982) provide detailed discussions of Bayesian methods. Anderson (1998) provides a discussion of Bayesian approaches based on frequency interpretations that may be more intuitive to those unfamiliar with probability interpretations.

A.5.0 REFERENCES

Anderson, J.L. 1998. "Embracing Uncertainty: The Interface of Bayesian Statistics and Cognitive Psychology." *Conservation Ecology* Vol. 2, Iss. 1, Art. 2.

Ang A. H-S and W.H. Tang. 1975. *Probability Concepts in Engineering Planning and Design, Vol. I*. John Wiley & Sons, New York, N.Y.

———. 1984. *Probability Concepts in Engineering Planning and Design, Vol. II*. John Wiley & Sons, New York, N.Y.

Benjamin, J.R. and A.C. Cornell. 1970. *Probability, Statistics, and Decision for Civil Engineers*. McGraw-Hill, New York, N.Y.

Chang, H.H. 1998. *Fluvial Processes in River Engineering*. Krieger, Malabar, FL.

Ecology (Washington Department of Ecology). 1993. "Supplement to Statistical Guidance for Ecology Site Managers." Memorandum dated August 11, 1983.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Von Nostrand Reinhold, New York, NY.

Hahn G.J., and S.S. Shaprio. 1967. *Statistical Models In Engineering*. John Wiley & Sons, New York, NY.

Hanover, W.J. 1980. *Practical Non-Parametric Statistics*. 2d ed. John Wiley & Sons, New York, N.Y.

Harr, M.E. 1987. *Reliability-Based Design in Civil Engineering*. McGraw-Hill, New York, N.Y.

Jornel, A.G. and CH. J. Huijbregts. 1978. *Mining Geostatistics*. Academic Press, New York, N.Y.

Kaplan, W. 1984. *Advanced Calculus.* 3d ed. Addison-Wesley, Reading, Mass.

Krumblein, W.C., and F.A. Graybill. 1965. *An Introduction to Statistical Models in Geology*, McGraw-Hill, New York, N.Y.

Martz, H.F. and R.A. Waller. 1982. *Bayesian Reliability Analysis*. John Wiley & Sons, New York, N.Y.

Meyer, S. L. 1975. *Data Analysis for Scientists and Engineers*. John Wiley & Sons, New York, N.Y.

Natrella, M.G. 1966. *Experimental Statistics*. National Bureau of Standards Handbook 91. USGPO, Washington, D.C.

Ott, W.R. 1995. *Environmental Statistics and Data Analysis*. CRC Lewis Publishers, New York, N.Y.

Raiffa H. and R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Harvard Business School, Mass.

U.S. Environmental Protection Agency (USEPA). 2000a. *Coeur d'Alene Basin RI/FS, Remedial Investigation Report.* Draft (Revision 1). Prepared by URSG and CH2M HILL for USEPA Region 10, under ARCS Contract No. 68-W9-0054. Seattle, WA. October 2000.

———. 2000b. *Coeur d'Alene Basin RI/FS, Feasibility Study Report.* Draft (Revision 1). Prepared by URSG and CH2M HILL for USEPA Region 10, under ARCS Contract No. 68-W9-0054. Seattle, WA. December 2000.

———. 1997. *Smelterville Flats Tailings Removal Plan, Draft Technical Memorandum, Task 16.03, Bunker Hill Superfund Site, Kellogg, Idaho*. Prepared by CH2M HILL for USEPA under ARCS Contract No. 68-W9-0031. February 1997.

———. 1992a. *Statistical Methods for Evaluating the Attainment of Cleanup Standards. Volume 3: Referenced Based Standards For Soils and Solid Media*. EPA 230-R-94-004. December 1992.

———. 1992b. *Methods for Evaluating the Attainment of Cleanup Standards. Volume 2: Ground Water*. EPA 230-R-92-0xx. July 1992.

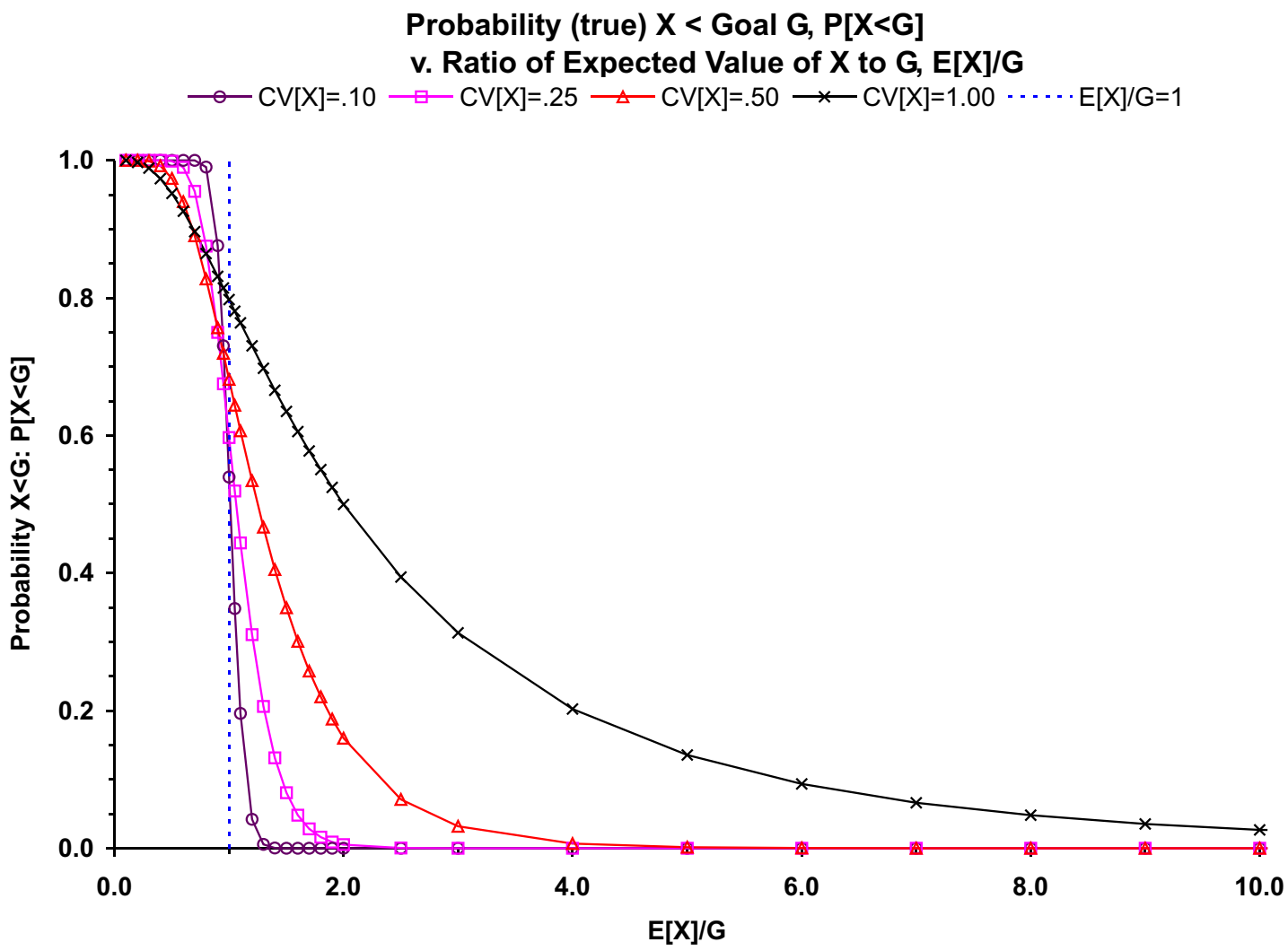
———. 1989. *Methods for Evaluating the Attainment of Cleanup Standards. Volume 1: Soils and Solid Media*. EPA 230/02-89-042. February 1989.

U.S. Navy. 2001. *Chemical Fate and Transport, Bremerton Naval Complex, Terrestrial OU B, Draft Technical Memorandum*. Prepared by URS Corp. for Engineering Field Activity Northwest, under CLEAN Contract #N62474-89-D-9295. Poulsbo, WA. May 31, 2001.

———. 1997. "Fate and Transport Methodology," *Focused Feasibility Study for Petroleum Sites, Adak Naval Complex, Adak Island, Alaska (Revision 1).* Prepared by URS Corp. for Engineering Field Activity Northwest, under CLEAN Contract #N62474-89-D-9295. Poulsbo, WA. Oct. 3, 1997.

Vanmarcke, E. 1983. *Random Fields: Analysis and Synthesis*. MIT Press, Cambridge, Mass.

Vogel, Richard M., Beth E. Rudolph, and Richard P. Hooper. 2005. Probabilistic Behavior of Water-Quality Loads. *Journal of Environmental Engineering*. ASCE. July.

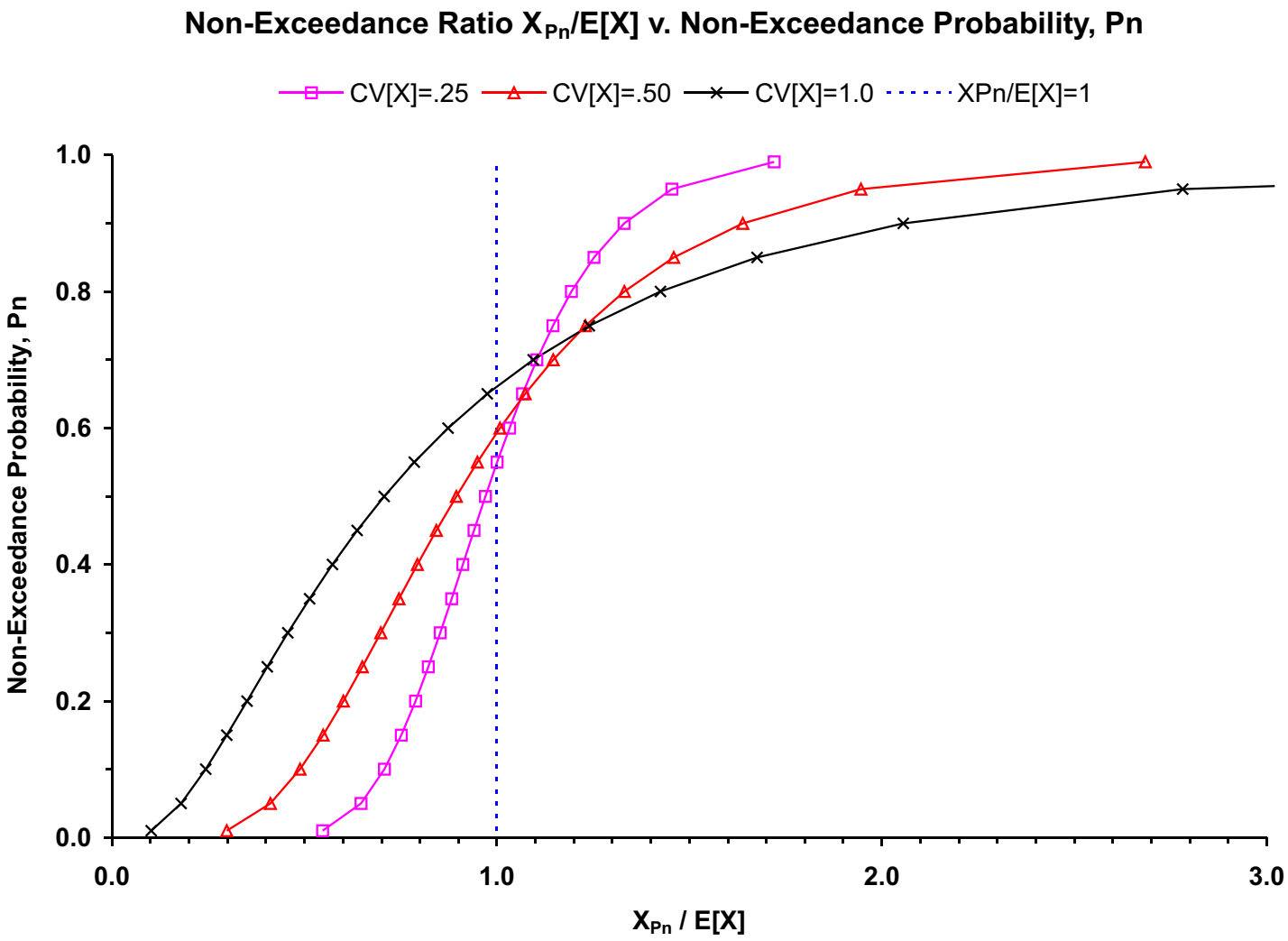


EPA REGION 10

Coeur D'Alene Basin RI/FS
A PREDICATIVE ANALYSIS FOR
POST-REMEDIATION METAL LOADING
Appendix A
Doc Control: 33753909-7326

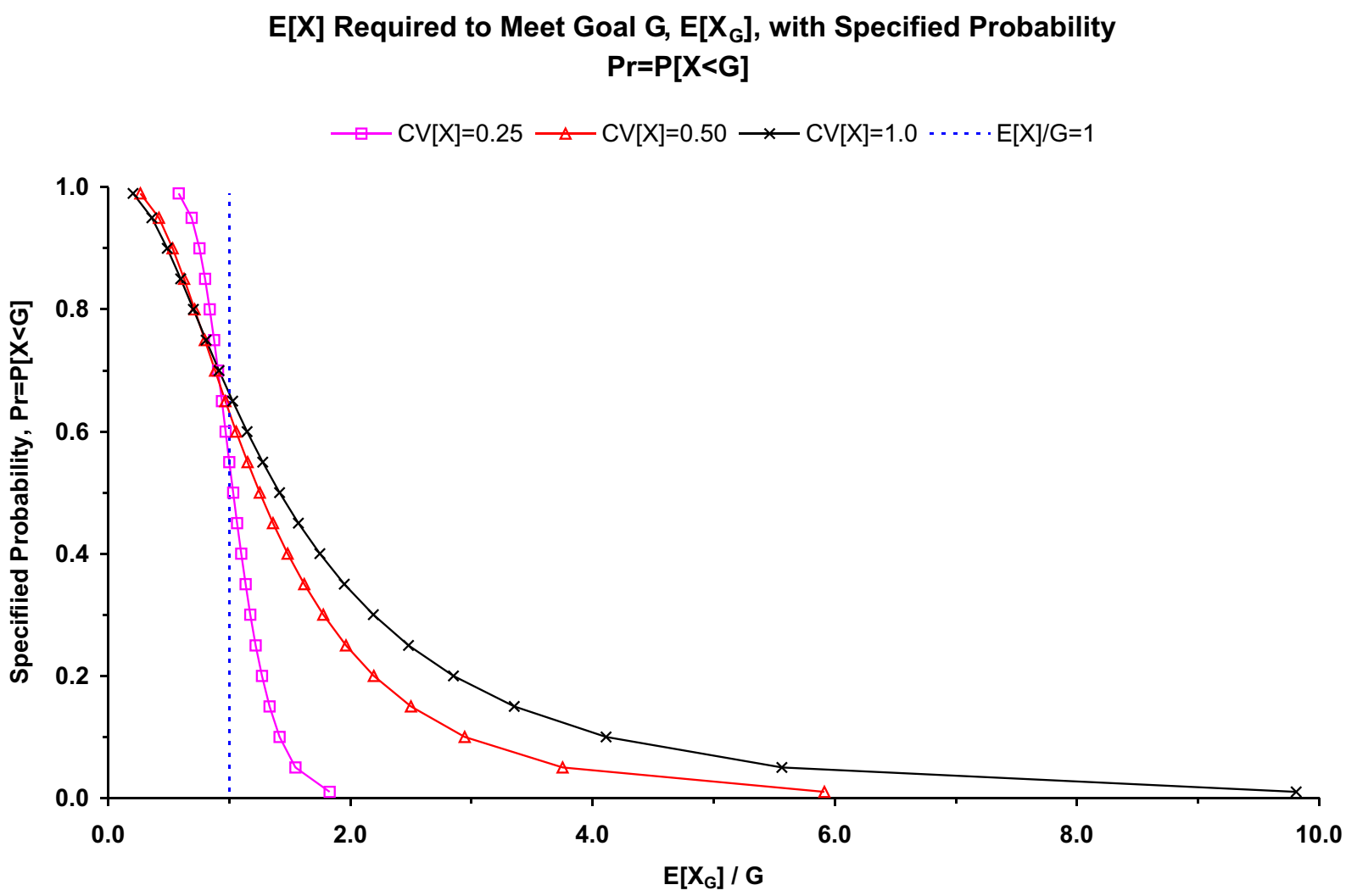
Ref: 027-RI-CO-102
Coeur D'Alene Basin RI/FS
Doc Control: 4162500.6538.05a
Probabilistic Series
09/13/01

Figure A-1



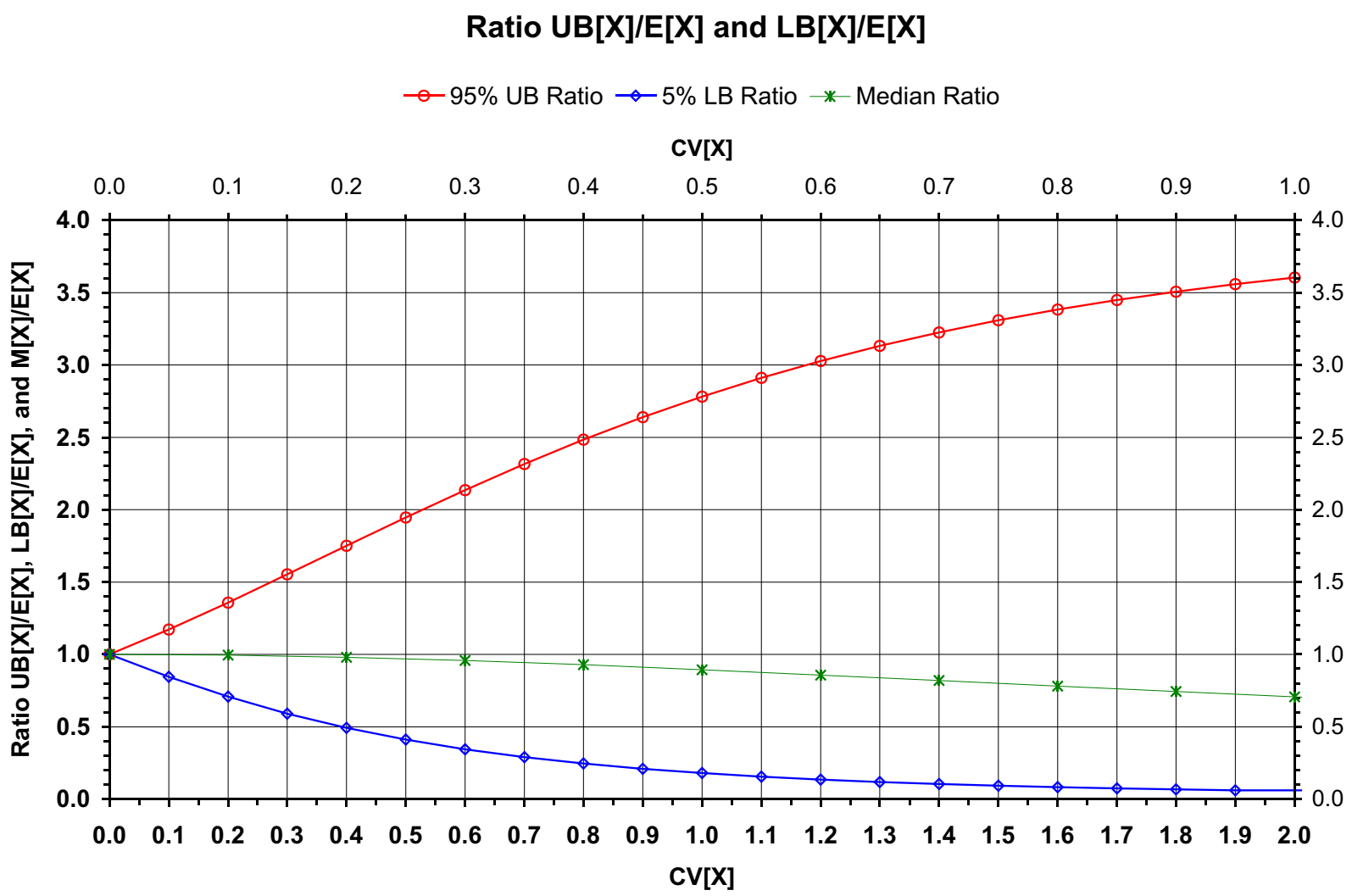
EPA REGION 10	Coeur D'Alene Basin RI/FS A PREDICATIVE ANALYSIS FOR POST-REMEDIATION METAL LOADING Appendix A Doc Control: 33753909-7326	Ref: 027-RI-CO-102 Coeur D'Alene Basin RI/FS Doc Control: 4162500.6538.05a Probabilistic Series 09/13/01

Figure A-2



EPA REGION 10	Coeur D'Alene Basin RI/FS A PREDICATIVE ANALYSIS FOR POST-REMEDIATION METAL LOADING Appendix A Doc Control: 33753909-7326	Ref: 027-RI-CO-102 Coeur D'Alene Basin RI/FS Doc Control: 4162500.6538.05a Probabilistic Series 09/13/01

Figure A-3



EPA REGION 10

Coeur D'Alene Basin RI/FS
A PREDICATIVE ANALYSIS FOR
POST-REMEDIATION METAL LOADING
Appendix A
Doc Control: 33753909-7326

Ref: 027-RI-CO-102
Coeur D'Alene Basin RI/FS
Doc Control: 4162500.6538.05a
Probabilistic Series
09/13/01

Figure A-4