Appendices to the Occurrence and Exposure Assessment for the Final Long Term 2 Enhanced Surface Water Treatment Rule

Office of Water (4606-M) EPA 815-R-06-002 December 2005 www.epa.gov/safewater

Appendix A. Waterborne Outbreaks Caused by Microbial Agents in Public Water Systems 1991-2000

Year	State	Cases	Etiology	System	Deficiency	Location	Source
1991	CA	15	*Giardia*	NC	4	Recreation area	Spring
1991	IL	386	AGI	NC	5	School	Well
1991	MI	1,320	AGI	NC	2	Campground	Well
1991	MI	33	AGI	NC	2	Resort	Well
1991	MN	30	AGI	NC	2	Campground	Well
1991	MN	30	AGI	NC	4	Resort	Well
1991	MN	17	AGI	NC	2	Restaurant	Well
1991	NM	38	AGI	NC	2	Camp	Well
1991	PA	170	AGI	NC	3	Picnic Area	Well
1991	PA	8	AGI	NC	3	Restaurant	Well
1991	PA	13	*Giardia*	NC	3	Park	Well
1991	PA	551	*Cryptosporidium*	NC	3	Picnic Area	Well
1991	PA	300	AGI	NC	3	Camp	Well
1991	PR	202	AGI	C	4	Penitentiary	River
1991	PR	9,847	AGI	C	3	Community	River
1992	ID	15	*Giardia*	C	2	Trailer Park	Well
1992	MN	250	AGI	NC	3	Restaurant	Lake
1992	NV	80	*Giardia*	C	3	Community	Lake
1992	NY	107	AGI	NC	4	Restaurant	Well
1992	NC	200	AGI	NC	2	Restaurant	Well
1992	OH	129	AGI	NC	4	Campground	Well
1992	OR	3,000	*Cryptosporidium*	C	3	Community	Spring
1992	OR		*Cryptosporidium*	C	3	Community	River
1992	PA	5	AGI	NC	3	Restaurant	Well
1992	PA	28	AGI	C	5	Park	River
1992	PA	42	AGI	NC	3	Camp	Well
1992	PA	50	AGI	NC	3	Camp	Well
1992	PA	57	AGI	NC	3	Camp	Well
1992	PA	80	AGI	NC	3	Camp	Well
1992	WY	150	*Shigella sonnei*	NC	2	Park	Well
1993	MN	27	*Cryptosporidium*	NC	5	Resort	Lake
1993	MO	625	Salmonella serotype Typhimurium	C	4	Community	Well
1993	NV	103	*Cryptosporidium*	C	5	Community	Lake
1993	NY	172	*Campylobacter jejuni*	C	5	Subdivision	Well
1993	PA	20	*Giardia lamblia*	C	3	Trailer Park	Well
1993	PA	65	AGI	NC	3	Ski Resort	Well
1993	SD	7	*Giardia*	C	2	Subdivision	Well
1993	SD	40	AGI	NC	2	Resort	Well
1993	WI	403,000	*Cryptosporidium*	C	3	Community	Lake
1994	IN	118	AGI	NC	2	Restaurant	Well
1994	ME	72	AGI	NC	2	Camp	Well
1994	MN	19	*Campylobacter jejuni*	NC	2	Park	Well
1994	NH	18	*Giardia*	C	3	Community	Reservoir
1994	NH	36	*Giardia*	C	3	Community	Lake

Year	State	Cases	Etiology	System	Deficiency	Location	Source
1994	NY	230	*Shigella sonnei*	NC	2	Camp	Well
1994	Saipan	11	Non-01 *Vibrio cholerae*	C	5	Bottled Water	Wells
1994	PA	200	AGI	NC	3	Resort	Well
1994	TN	304	*Giardia*	C	4	Correctional Facility	Reservoir
1994	WA	134	*Cryptosporidium*	C	2	Community	Well
1995	ID	83	*Shigella sonnei*	NC	2	Resort	Well
1995	ID	18	AGI	C	3	Community	Well
1995	MN	33	*E. coli* O157:H7	NC	3	Camp	Spring
1995	MT	450	AGI	NC	2	Campground	Well
1995	NY	1,449	*Giardia*	C	3	Water utility	Lake
1995	OK	10	*Shigella sonnei*	NC	3	Store	Well
1995	PA	19	AGI	NC	2	Inn	Well
1995	SD	48	AGI	NC	2	Camp	Well
1995	WA	87	*Giardia*	C	4	Community	Well
1995	WI	26	AGI	NC	3	Restaurant	Well
1995	WI	148	Small round structured virus	C	4	School	Lake
1996	ID	94	AGI	NC	3	Camp	Well
1996	NY	60	*Plesiomonas shigelloides*	NC	3	Restaurant	Spring
1996	WI	21	AGI	NC	4	Restaurant	Well
1997	CO	9	AGI	NC	3	Cabins	Spring
1997	NM	123	AGI	NC	4	Country club	Well
1997	NY	1,450	Norwalk-like virus	NC	3	Ski resort	Well
1997	NY	50	*Giardia*	C	3	Community	Lake
1997	OR	100	*Giardia*	NC	4	Campground	Well/Spring
1997	SD	16	AGI	NC	3	Campground	Well
1997	WA	4	*E. coli* O157:H7	NC	3	Trailer Park	Well
1998	FL	7	*Giardia*	C	2	Community	Well
1998	MN	83	*Shigella sonnei*	C	4	Fairgrounds	Well
1998	OH	10	AGI	C	4	Treatment plant	Surface
1998	TX	1,400	*Cryptosporidium*	C	3	Subdivision	Well
1998	WY	157	*E. coli* O157:H7	C	2	Community	Well/Spring
1999	CA	31	AGI	NC	2	Camp	Well
1999	FL	4	AGI	C	2	Community	Well
1999	FL	6	AGI	C	4	Apartment	River/stream
1999	FL	3	AGI	C	4	Community	Well
1999	MO	124	*Salmonella typhimurium*	C	3	Community	Well
1999	NM	70	Small round-structured virus	NC	3	Camp	Spring
1999	NY	781	*E. coli* O157:H7/ *Campylobacter jejuni*	NC	2	Fairground	Well
1999	TX	22	*E. coli* O157:H7	C	3	Community	Well
1999	WA	68	AGI	NC	2	Soccer match	Well
2000	CA	147	Norwalk-like virus	NC	2	Camp	Well
2000	CO	27	*Giardia*	NC	3	Resort	River
2000	FL	19	AGI	C	3	Trailer park	Well
2000	FL	21	AGI	C	3	Trailer park	Well
2000	FL	5	*Cryptosporidium*	C	4	Community	Well

Year	State	Cases	Etiology	System	Deficiency	Location	Source
2000	ID	15	*Campylobacter jejuni*	NC	2	Camp	Spring
2000	ID	65	AGI	NC	2	Restaurant	Well
2000	KS	86	Norwalk-like virus	NC	2	Reception hall	Well
2000	MN	12	*Giardia*	NC	2	Camp	Well
2000	OH	29	*E. coli* O157:H7	C	4	Fairground	Surface
2000	WV	123	Norwalk-like virus	NC	3	Camp	Well

AGI = Acute gastrointestinal illness of unknown etiology.
NC = Non-community; C = community
Definitions of deficiencies: (1) Untreated surface water; (2) untreated ground water; (3) treatment deficiency (e.g., temporary interruption of disinfection, chronically inadequate disinfection, and inadequate or no filtration); (4) distribution system deficiency (e.g., cross-connection, contamination of water mains during construction or repair, and contamination of a storage facility); and (5) unknown or miscellaneous deficiency (e.g., contaminated bottled water.

Sources: Moore et al. 1993, Kramer et al. 1996, Levy et al. 1998, Barwick et al. 2000, and Lee et al. 2002

Appendix B. Modeling Microbial Source Water Occurrence

Chapter 3 of this document discusses the primary sources of measurement data used by EPA to characterize the occurrence of *Cryptosporidium* and other pathogens in surface water used as drinking water sources. That discussion addresses the methods that were used to collect and analyze those data. It also addresses the statistical models that EPA developed to use in conjunction with those measurement data to derive a plausible range of estimates of the actual (but unknown) national distribution of *Cryptosporidium* occurrence in the source waters used by public water supplies. Modeling is necessary because the national occurrence of *Cryptosporidium* cannot be fully revealed by the available measurement data alone due to a variety of limitations and uncertainties inherent in those measurements.

Chapter 4 of this document presents both descriptive summaries of the pathogen occurrence measurement data collected by EPA, and the results of the statistical modeling performed by EPA to characterize national occurrence based upon the measurement data.

The purpose of this Appendix is to provide additional technical detail on EPA's approach to the statistical modeling discussed in Chapter 3 and that was used to derive the occurrence information presented in Chapter 4.

This appendix also presents several evaluations done by EPA to examine the validity and implications of some of the key assumptions made in the modeling that was performed.

This appendix is organized into the following major sections:

B.1 Model Overview
B.2 Model Structure
B.3 Model Inputs
B.4 Model Fitting and Outputs
B.5 Model Evaluations
B.6 Reduced-Form Model

As indicated in Chapter 3, EPA initially developed the full form of the microbial occurrence model, which is described in sections B.1 through B.6, for filtered plants which comprise the majority of all surface water systems in the U.S. EPA also developed a reduced form of the model for unfiltered plants primarily to better accommodate the more limited input data available for those plants compared to the filtered plants. For consistency sake in implementing these models for evaluating the impacts of regulatory alternatives, EPA chose to also develop a reduced form of the model for filtered plants as well. The reduced-form model that was used for the economic impact analysis is described in Section B.6.

B.1 Model Overview

There are several related objectives of the *Cryptosporidium* occurrence modeling performed by EPA. Key among those objectives is to provide a scientifically defensible characterization of the national distribution of this pathogen in surface waters that are used as a source by public drinking water systems. This information is critical for understanding the current risks of endemic cryptosporidiosis among those served by surface water systems, and to estimate and compare the costs and benefits of reducing that risk from the implementation of treatment changes to comply with several regulatory alternatives being considered by EPA for the LT2ESWTR. In addition to that overarching objective, the occurrence modeling effort also identifies important relationships between pathogen occurrence and other specific characteristics of the source waters examined to help guide the development of regulatory and treatment alternatives to most effectively eliminate or minimize the risks posed by *Cryptosporidium* for public water supplies.

There are two main facets of the *Cryptosporidium* occurrence modeling. The first focuses on modeling the expected average concentration of *Cryptosporidium* at individual locations reflecting the observed data, uncertainties and limitations of the sampling and analysis procedures, and possible influence of other characteristics of the source water being considered. The second facet of the modeling focuses on characterizing the national distribution of the average levels of *Cryptosporidium* in source waters used by public water systems based upon the aggregation of modeled levels at specific locations obtained in the first stage of the modeling.

The modeling performed to characterize average *Cryptosporidium* at individual locations involves the basic assumption that measurements taken at a individual locations will each follow a Poisson distribution specific to that site. As discussed below, there are a number of general and specific technical issues that must be addressed in deriving the specific Poisson distribution for each location.

The modeling performed to characterize the national distribution of average *Cryptosporidium* concentrations involves the assumption that the distribution of those individual location averages can be characterized by a lognormal distribution.

Another important consideration in modeling both the individual location averages and the national distribution of those individual location averages is the recognition of the many limitations and uncertainties in the underlying measurement data, as well as those resulting from other modeling assumptions used by EPA. Therefore, another key objective of the modeling approach used by EPA was to capture and reflect that uncertainty in the model outputs. As described more fully in the sections that follow, the results of this modeling are not limited to a “best estimate” of occurrence, but rather these results are presented as sets of plausible occurrence distributions that are consistent with the underlying observations.

B.1.1 Overview of Modeling Occurrence at a Single Location

As discussed by Haas et al. (1999), the most appropriate probability distribution for characterizing the occurrence of microorganisms in source water at a particular location is the *Poisson distribution*. The Poisson distribution is a fundamental probability distribution that is applied when the average number of occurrences of a discrete event is the result of a large collection of situations in which that event could occur, and a very small probability for it to occur in any one specific situation. It is used extensively to address problems that arise in the counting of relatively rare and independent events occurring in some unit interval of time, length, area, or volume (Sachs 1984).

Some examples of discrete, independent events occurring in some unit interval that may be appropriately described by the Poisson distribution are radioactive disintegration (time interval), material irregularities in a wire or surface (length or area interval), and raisins in raisin bread (volume interval).

The Poisson distribution is commonly expressed mathematically as:

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

This equation describes the probability that in some randomly selected interval (or volume or area) the observed number of events *X* will equal some specified value *x*, given that the known average (or expected) number of events is equal to the value λ. The only parameter of the Poisson distribution is λ, the average (or expected) number of events in that unit interval.

The occurrence of microorganisms in a unit volume of water is also a type of phenomenon that can be appropriately described by the Poisson distribution. EPA has used the *Cryptosporidium* monitoring data obtained from the Information Collection Rule (ICR) monitoring program, the ICR Supplemental Surveys, and the Poisson distribution assumption to derive estimates of the average *Cryptosporidium* concentrations in the source waters used by the plants included in those surveys. More specifically, EPA has used these monitoring data and model assumptions to derive a range of plausible estimates of the underlying source water occurrence of *Cryptosporidium* that are consistent with the observations and that also reflect the limitations and uncertainties inherent in collecting and analyzing those data.

A basic type of question about *Cryptosporidium* occurrence in the source water used by an individual plant that the Poisson distribution is suited to answer would be the following:

> If the actual average concentration of *Cryptosporidium* in the source water is 1 oocysts per liter, what is the probability that an analyst examining a randomly selected one liter volume of water will observe exactly one oocyst? Or two oocysts? Or zero oocyts?

If, as stated above, the known underlying concentration of *Cryptosporidium* is 1 oocyst per liter, then one would compute from the Poisson distribution that the probability of observing exactly 1 oocyst in a randomly collected 1-liter sample is the following:

$$P(X = 1|1) = \frac{1^1 e^{-1}}{1!} = e^{-1} = 0.368$$

That is, given that the actual average *Cryptosporidium* concentration is 1 oocyst per liter, it would be expected that one will actually observe exactly 1 oocyst in a randomly selected 1-liter sample only about 37 percent of the time.

Similarly, the expectation of observing zero oocysts in a liter sample given a 1 oocyst per liter average is the following:

$$P(X = 0|1) = \frac{1^0 e^{-\lambda}}{0!} = e^{-1} = 0.368$$

That is, under these circumstances it is just as likely to observe zero oocysts in a random 1-liter sample as it is to observe 1 oocyst in that sample, even when the known underlying concentration is 1 oocyst per liter. Note that under these assumptions, there is an approximately 18 percent chance of observing exactly 2 oocysts, and approximately 6 percent chance of observing 3 oocysts, leaving about a 2 percent chance of observing 4 or more oocysts in any randomly selected 1-liter sample.

In the above example, the assumed actual concentration of *Cryptosporidium* of 1 oocyst per liter is the upper end of the range expected to be encountered in source waters. A more typical source water concentration would be 0.1 oocyst per liter. In the ICR, the median sample volume analyzed was approximately 3 liters. With a concentration of 0.1 oocycst per liter and a 3-liter sample, the expected count (λ) is 0.3 (=0.1 * 3). Using the Poisson distribution, the probability of seeing zero oocysts in a randomly drawn 3-liter sample is the following:

$$P(X = 0|0.3) = \frac{0.3^0 \cdot e^{-0.3}}{0!} = e^{-0.3} = 0.741$$

That is, even though *Cryptosporidium* is present at a concentration of 0.1 oocyst per liter, just over 74 percent of all randomly selected 3-liter samples are expected to result in observations (measurements) of zero. The expectation of a observing substantial number of zero count measurements even when *Cryptosporidium* is present in the source water is an important factor to keep in mind when considering how the Poisson model is actually used in the occurrence modeling.

The use of the Poisson distribution as shown in the above examples allows one to calculate the probability of observing a particular number of events (in this case, counts of *Cryptosporidium* oocysts present in some unit volume of water) when one already knows the model's parameter λ (in this case, the expected count of *Cryptosporidium* in that unit volume of source water—which is to say, the average concentration of *Cryptosporidium* in association with some specified sample volume). However, what is actually needed by EPA are estimates of the average *Cryptosporidium* concentrations in the sampled source waters that have resulted in the observed occurrence data at various locations studied in the ICR and ICRSS. Therefore, the modeling effort undertaken by EPA in this first facet of the overall modeling process is focused on estimating the Poisson distribution λ parameters—or more specifically, the underlying concentrations reflected in those parameters—based upon the observed measurement data from the ICR and ICRSS.

The process of estimating the λ parameter for a Poisson distribution can often be a challenge because one of the characteristics that gives rise to the Poisson distribution is that the events being characterized are relatively rare. Ideally, one would be able to make a large number of reliable and representative observations so that a sufficient number of the rare events are observed in order to estimate the value of the λ parameter with a high degree of confidence. In many cases, however, there are limitations on the number of observations that can be made and uncertainties inherent in the collection and measurement of the data that are used to derive the model parameter.

Several such difficulties occur in the case of *Cryptosporidium* measurements in the ICR and ICRSS data. Key among the difficulties encountered are the relatively small (and nonuniform) sample volumes collected in those studies, the limited number of total samples taken, and measurement difficulties that result in less than 100 percent recovery (counting) of all of the oocysts that are actually present in a sample. These difficulties, and the efforts taken by EPA to overcome them, are discussed in the next several sections of this appendix.

B.1.2 Overview of Modeling National Occurrence

Because the ICR and ICRSS do not provide data for *Cryptosporidium* occurrence at all times for all source waters used by public water systems, it is necessary to consider the occurrence data and modeling results for the individual locations included in those surveys as representative samples.

EPA has used the basic assumption that the national distribution of plant-mean *Cryptosporidium* levels can be modeled as a lognormal distribution. The lognormal distribution is another fundamental probability distribution that is used commonly and effectively to characterize environmental contaminant occurrence. The basic characteristic of a lognormal distribution is that the logarithms of the values being evaluated (in this case, the plant-mean concentrations of *Cryptosporidium* in source waters) are normally distributed. One property of the lognormal distribution that makes it particularly well-suited to describing phenomena like environmental contaminant occurrence data is that it is bounded by zero on the low end and it reflects a "right-skewed" distribution—that is, it has a tail in the upper end—that is consistent with having a small proportion of values with relatively high values. The lognormal distribution has two parameters, the log mean (usually referred to as mu or μ) and the log standard deviation (usually referred to as sigma or σ).

In addition to a set of estimates of the plant-mean *Cryptosporidium* levels at the various sample locations from the first aspect of the modeling, the model also derives the parameters of a lognormal distribution that is consistent with those modeled plant-mean values. In the overall modeling framework, this process is repeated a large number of times to capture the uncertainty in both the estimates of the plant-means at individual locations and the uncertainty associated with estimating the lognormal distribution parameters. However, this process does not capture the following types of uncertainty: model uncertainty, uncertainty in the analytical method, and uncertainty associated with assuming that small systems are like medium and large systems. Note, as with all modeling efforts of this type, the scope of the uncertainty analysis is constrained by the specific distributional assumptions adopted in performing the hierarchical modeling, and therefore results obtained from the analysis represent a lower bound on the overall uncertainty.

B.1.3 Basis for Modeling

There are two features of the underlying data that suggest that modeling is an appropriate approach to estimating national occurrence. These are small sample volumes and low recovery rates, both of which operate to produce low counts of *Cryptosporidium* oocysts.

The first of these, small sample volumes, was touched on somewhat in the general description given above of the difficulties in parameterizing a Poisson distribution model. As noted there, the median sample volume size in the ICR was only 3 liters, and if the "true" underlying average concentration in the source water is 0.1 oocyst per liter, it is expected from the Poisson distribution that no *Cryptosporidium* oocysts would be observed in approximately 74 percent of the samples.

The second aspect of the measurement process is that the recovery rate for the methods was less than 100 percent. As discussed in Chapter 3, the mean recovery for the ICR method was 11.6 percent and for the ICRSS method was 43 percent. Therefore, this suggests that most of the oocysts present in the source water samples collected in these surveys many not have been counted in the assays performed.

A simple simulation analysis was performed to show the potential combined effect of both a relatively low sample volumes and the low recovery rate in the ICR across a range of possible "true"*Cryptosporidium* concentrations in a source water. Based on analyses of spiked samples using the

ICR methods, the recovery rate for the *Cryptosporidium* was modeled as a beta distribution with parameters α=1.44, β=11.20 (mean = 0.114 or 11.4 percent recovery) (Messner, 2000). Note that the beta distribution is the natural distribution for describing a continuous random variable with a value between zero and one. This implies that, on average, an oocyst actually present in a water sample would only be counted as such about 11 percent of the time. Note that the mean of the beta distribution, 11.4, differs slightly from the mean of 11.6 based on the spiked study.

Exhibit B.1a shows the distribution of recovery rates as the density function of a beta distribution having the parameters α=1.44, β=11.20.

Exhibit B.1a
Recovery Rate Distribution Described by a Beta Density Function with Parameters α=1.44, β=11.20 (average = 0.114)

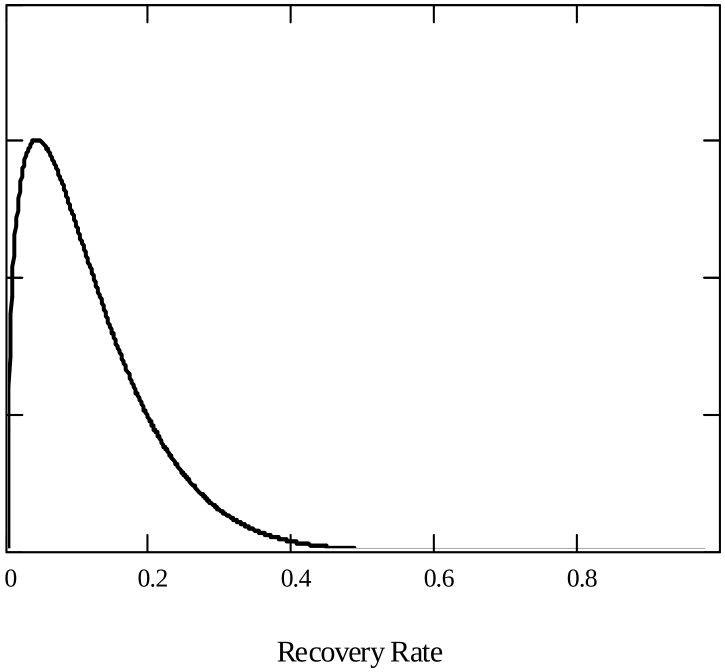
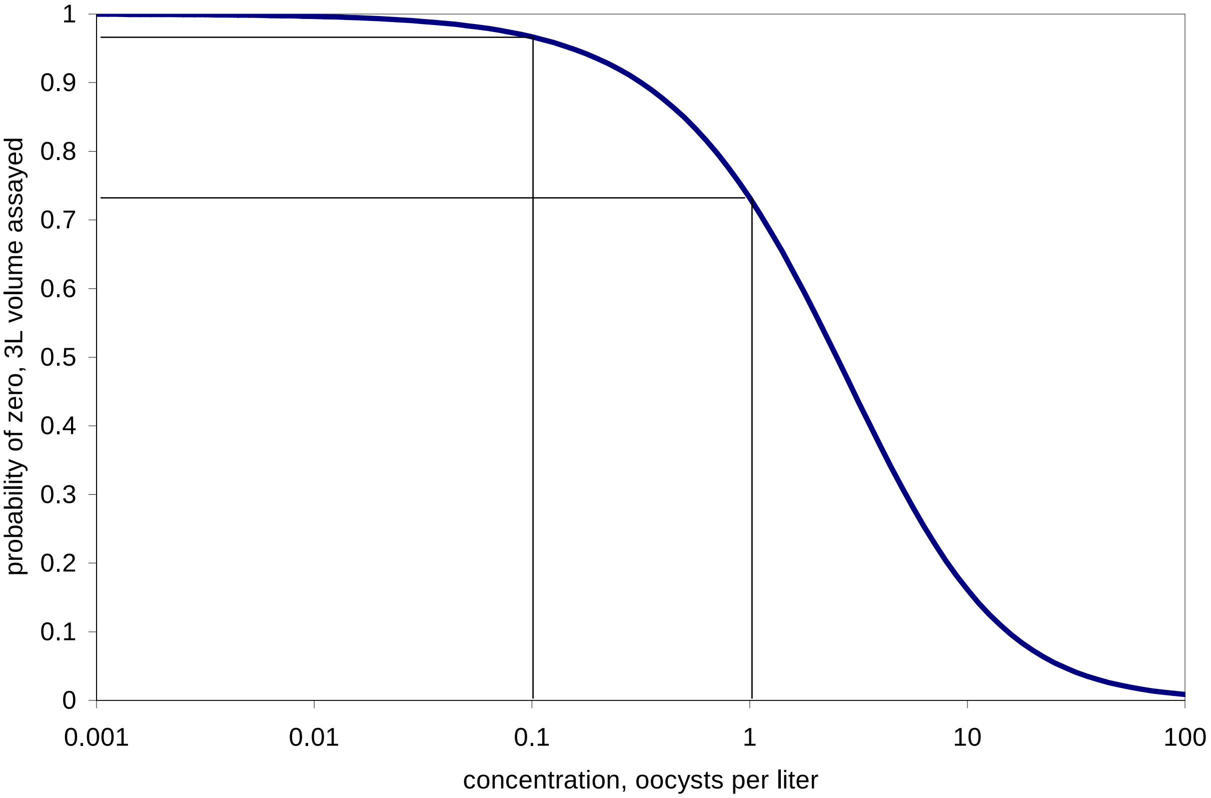


Exhibit B.1b depicts the combined effect of small volume assayed and low, variable recovery across a range of possible "true" source water concentrations of *Cryptosporidium*. These possible "true" average *Cryptosporidium* concentrations in the sampled water are shown on the x-axis of Exhibit B.1. The probability of observing zero oocysts in a 3-liter sample drawn from a source water having a specific underyling actual concentration is shown on the y-axis.

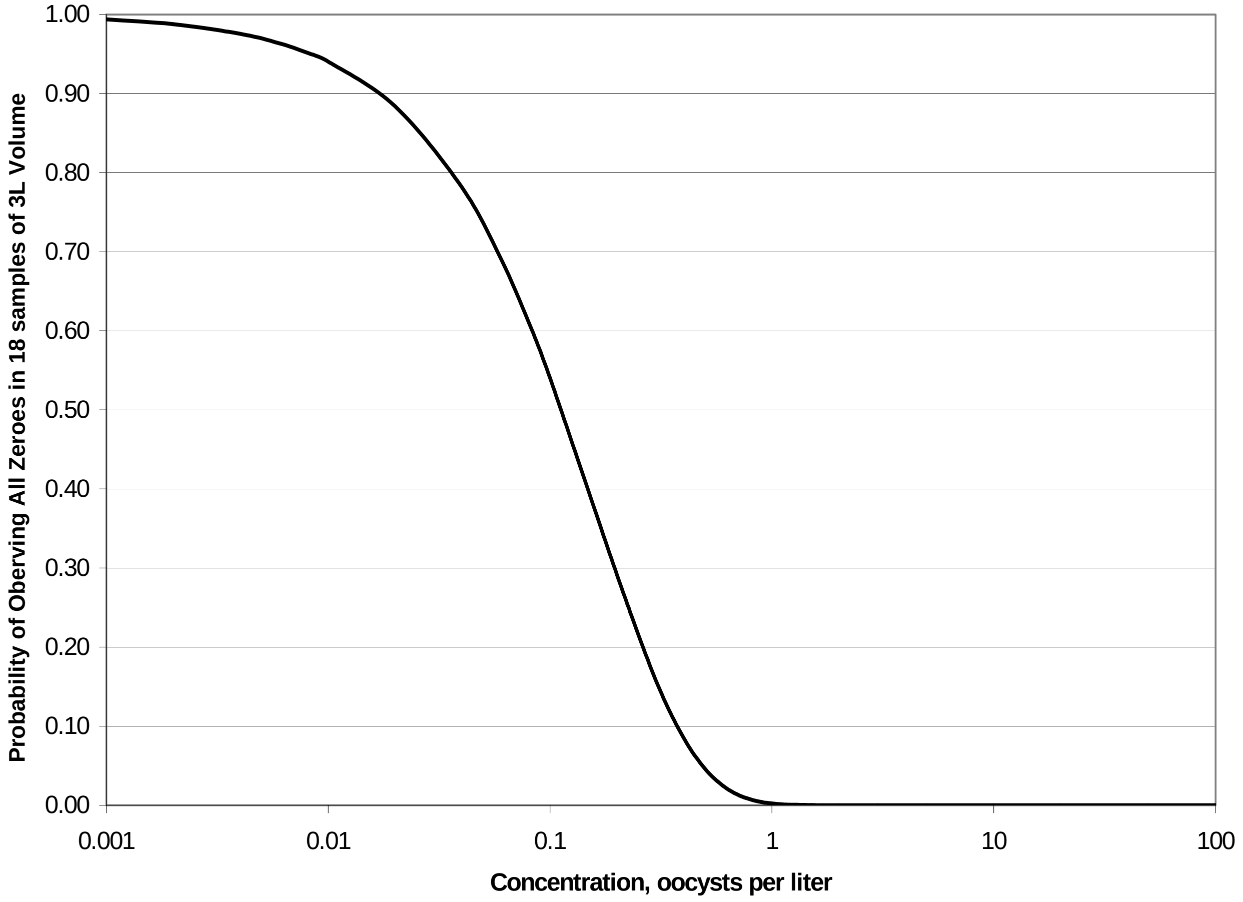
Exhibit B.1b
Expected Probability of Observing a Zero Count as a Function of Actual Oocyst Concentration, When 3 Liter Sample is Assayed and Recovery is Beta (1.44, 11.2)



As indicated by the two examples featured by the vertical and horizontal lines on the graph, a 3-liter sample drawn from a source water having 0.1 oocysts per liter, and a recovery rate varying about a mean of 11.4%, is expected to yield a zero count with probability 0.97. Even when the actual concentration is an order of magnitude higher at 1 oocyst per liter, the Exhibit shows that a 3-liter sample will yield a zero with probability 0.73.

An alternative view of this is provided in Exhibit B.1c. This graph displays the probability of observing all zeroes in 18 samples of 3 L each given the 'true' concentration shown on the x-axis and an average recovery rate of 11.4%. For example, if the concentration is 0.01 oocysts per L, then the probability of observing zero oocysts in a single sample is 1-exp[(-(3)(0.01)(0.114)] = 0.0034. The probability of observing zeroes in all 18 samples is $(1\text{-}0.0034)^{18}$ = 0.94.

Exhibit B.1c Probability of Observing All Zeroes in 18 Samples of 3 L Each for the Given Oocyst Concentration and Assuming 11.4% Average Recovery Rate



As noted previously, EPA developed the occurrence model to accommodate a number of limitations and uncertainties, including the effect of small sample volumes and low recovery rates, in order to characterize a range of plausible actual average concentrations of *Cryptosporidium* in sampled source waters that are consistent with the relatively low incidence of oocysts observed in those surveys.

B.2 Model Structure

There are three levels to the occurrence model; these levels are depicted in Exhibit B.2. At the lowest level are the modeled features of individual measurements. These include the observed measurement results (microbial counts, volumes assayed, turbidity values, and source water type) and the unobserved true concentrations, measurement recoveries, and residuals (ε_{ij} = difference between model-predicted and true concentration). These low-level variables are indexed by both i (location) and j (month). The middle level of the model includes effects for locations, months, and source water types.

Each mid-level variable has only one index. The highest level of parameters includes an intercept term (β_0 = overall mean of log-concentrations), a turbidity effect, and four precisions that define how lower-level effects are distributed. These high-level parameters are global and require no indices.

Exhibit B.2 Components of the Three Primary Levels of the Occurrence Model

TOP LEVEL

- Precisions = ϕ^{MSW}, ϕ^{loc}, ϕ^{month}, ϕ^{resid}
- Turbidity Effect = β_{turb}
- Intercept = β_0

MIDDLE LEVEL

- Source Water Type Effects = β_{MSW} ~ Normal(mean 0, precision ϕ^{MSW})
- Locational Effects = ε_i ~ Normal(mean 0, precision ϕ^{loc})
- Month Effects = ε_j ~ Normal(mean 0, precision ϕ^{month})

BOTTOM LEVEL

- Unobserved Components
  - Microbial Concentration = C_{ij}
  - Measurement System Recovery = r_{ij}
  - Residual = ε_{ij} = $\ln(C_{ij})$ - model-predicted $\text{Concentration}_{ij}$
- Observed Components
  - Volume Assayed = V_{ij}
  - Turbidity = turb_{ij} (and standardized value t_{ij})
  - Source Water Type = MSW_{ij}
  - Microbial Counts = Y_{ij} ~ Poisson(C_{ij} * r_{ij} * V_{ij})

B.2.1 Expected Counts

The observed counts, $Y_{i,j}$, are modeled as Poisson random variables:

$$Y_{i,j} \sim \text{Poisson}(\lambda_{i,j})$$

Where: i = 1 to number of sample locations (350),
j = 1 to number of sample periods (18)

This part of the model can be described as a two-dimensional grid of sample locations (rows) and sample periods (columns), with a different expected count, $\lambda_{i,j}$ for each cell in the grid. Each expected count is modeled as a function of four parameters:

$$\lambda_{i,j} = (\text{Concentration}_{i,j} \times \text{Volume}_{i,j} \times \text{Recovery}_{i,j}) + \text{FP}$$

The three subscripted parameters vary with each individual test result. For a given sample tested from location i and time period j:

Source Water Concentration ($\textit{Concentration}_{i,j}$)—the actual (but unknown and therefore modeled) underlying concentration of microbes in the source from which the sample was drawn.

Sample Volume Analyzed ($\textit{Volume}_{i,j}$)—the sample volume analyzed.

Recovery Rate ($\textit{Recovery}_{i,j}$)—the predicted ratio of microbes detected to microbes present in the sample (as a percent).

The fourth parameter is the same for all samples:

False Positives (*FP*)—a model adjustment for the possibility of false positives, or detections of microbes that are not actually present in the sample (e.g., algae or other constituents in the water may be mistaken for a *Cryptosporidium* oocyst).

As noted previously, the $\textit{Concentration}_{i,j}$ parameters are unknown and estimating these from the data is the major focus of the model. The $\textit{Volume}_{i,j}$ values, on the other hand, are known. These are the sample volumes analyzed, along with sample counts, in the ICR and ICRSS. Because sample volume analyzed varied widely, and larger samples will contain more microbes, on average, for a given true concentration, these sample volumes are important predictors of the expected counts.

The $\textit{Recovery}_{i,j}$, values are not known, but are simulated based on results from test method evaluations. For each test method employed in the ICR and ICRSS, recovery was evaluated by testing “spiked” samples with known concentrations. These experiments allowed for direct estimation of the number of microbes detected versus the number actually present in the sample. From these experiments, a recovery rate distribution was estimated for each analytical method to capture the typical variation in recovery rates, and the modeled $\textit{Recovery}_{i,j}$ values are drawn randomly from these distributions.

The *FP* adjustment is derived from: 1) an assumed false positive rate, *fp*, for a given test method, and 2) the average number of microbes detected (positives) per sample in the data set. For example, if *fp* is assumed to be 1 percent, and a given data set shows a total of 100 microbes detected in 1000 samples, *FP* would be set equal to 0.001 (0.01 false positives per positive × 0.1 positives per sample = 0.001 false positives per sample).

Based on expert understanding of the analytical method results, reasonable false positive rates were tested during model development. The value used is shown in section B.3.3.

B.2.3 Modeling Contributions to Concentration

The modeling of expected counts provides a basic probability structure that links the primary data (sample volumes and laboratory counts of oocysts) to the values of primary interest—possible true source water microbial concentrations. In this next step, these possible true source water concentrations are further broken down as follows:

$$\text{Concentration}_{i,j} = R \times \exp(\beta_0 + \beta_1 t_{i,j} + \beta_{2:5}\text{MSW}_{i,j} + \varepsilon_i + \varepsilon_j + \varepsilon_{i,j})$$

Where:

- R = 0 for Z_i fraction of concentrations, and 1 for (1-Z_i) fraction of concentrations
- Z_i = Assumed true zero probability.

- β_0 = an intercept term that reflects overall log concentration across all i locations and j sample periods. Other parameters model deviations from this overall mean.

- β_1 = Regression parameter for turbidity.

- $t_{i,j}$ = Log_{10} of observed turbidity value (in nephelometric turbidity units (NTU), at location (i) and sample period (j). Measured turbidity values are standardized (re-scaled by adding a constant to have overall mean zero and standard deviation of one) before input to the model. This preserves $e\hat{}\beta_0$ as the natural log of the overall median concentration, and also allows for easy interpretation of the magnitude of β_1 estimates (relative to the other model parameter estimates, which are all on the microbe concentration scale, not the turbidity scale).

- $\text{MSW}_{i,j}$ = Type of source water (mixed surface water) – 1) surface flowing stream, 2) surface reservoir/lake, 3) ground water under the influence of surface water, 4) mixed surface and groundwater.

- $\beta_{2:5}$ = The MSW fixed effects. The β's allow each MSW class to have a different concentration.

- ε_i = The location random effect that allows each location to have a different concentration (it's a "random" rather than "fixed" effect because we are more interested in how these location effects are distributed than in any particular estimated ε_i).

- ε_j = Monthly random effect that allows each sample period to have a different concentration. An important distinction is that ε_i and ε_j are crossed, not nested, effects, which means that the ε_j measure monthly effects common to all locations and not just within a particular location.

- $\varepsilon_{i,j}$ = Residual term that embodies all other variation and uncertainty.

The true zero parameter accounts for the possibility that a particular water source is entirely free of a particular microbe. Since the exponential term in this equation for concentration is always greater than zero, the exponential term is multiplied by a 0/1 random variable, R, that is governed by a "true zero probability" model parameter. This parameter can be easily varied to explore model sensitivity to changes in the assumed "true zero" probability rate.

Note that when R is set equal to one (the usual case) the natural log of concentration is modeled as a linear function:

$$\ln(\text{Concentration}_{i,j}) = \beta_0 + \beta_1 t_{i,j} + \beta_{2:6}\text{MSW}_{i,j} + \varepsilon_i + \varepsilon_j + \varepsilon_{i,j}$$

Prior distributions are required for the unknown β parameters and for the variance (precision) parameters on the ε distributions since Bayesian techniques are used to estimate their true values. In these models, broad uncertainty is expressed by using widely dispersed prior distributions that allow the modeled results to rely largely on the data to drive the parameter estimates.

B.3 Model Inputs

Given the complexity of the occurrence model, it is easy to lose track of exactly where the data inputs end and the model assumptions begin. To reinforce these distinctions, this section summarizes the inputs to the *Cryptosporidium* occurrence models discussed. Much of this information has been discussed earlier in this appendix and also in Chapters 3 and 4 of the document. Rather than address it all in detail again, the goal here is to primarily list all the inputs concisely, in one place, and in a logical framework that clarifies how each contributes to the overall modeling.

B.3.1 Survey Data: Counts, Sample Volumes, Turbidity, and Source Water Type

There are six inputs that come directly from the ICR and ICR Supplemental Surveys. The following comprise the raw data inputs:

1) Microbial counts
2) Associated sample volumes
3) Associated turbidity measurement
4) MSW categorization (e.g., flowing river/stream, reservoir/lake)
5) Sample location
6) Sample month

B.3.2 Simulated Test Method Recovery Rate

This is a simulated, random input to each model. Recovery values are sampled from the following probability distributions:

ICR: beta distribution with parameters α=1.44, β=11.20 (mean = 0.114 or 11.4 percent)
ICRSS: beta distribution with parameters α=2, β=3 (mean = 0.400 or 40 percent)

The beta distribution generates values between zero and one. Here it used to characterize a range of recovery rates from zero (no oocysts ever detected, regardless of how many are in the test sample) to one (all oocysts in the sample are detected). Based on spiked sample evaluations, these beta distribution parameters were chosen to closely approximate, based on the best available estimates, the true range of recovery rates in actual *Cryptosporidium* testing (including sample to sample variation in this true rate).

For both the ICR and ICRSS, there are slight differences in the measured and modeled means (11.6 vs. 11.4 percent for the ICR and 43 vs. 40 percent for the ICRSS).

B.3.3 Tuneable Model Inputs: False Positives, True Zero

There are two "tuneable" inputs to the occurrence models: 1) the false positive rate, and 2) the true proportion of systems with source water that is completely free of the target microbe. These model parameters could be easily changed in the model development process to both reflect expert opinion and to assess the impact of changing the parameter on overall model results.

False positive rates were based largely on expert opinion. In both ICR and ICRSS modeling, a false positive rate of 0.01 was assumed for total *Cryptosporidium* counts (the category of count summarized in Chapter 4 modeled distribution curves for plant-mean concentration).

In developing the model, several values for true zero were tested, ranging from 0 to 50 percent. Experts believed true zero concentrations rarely occur and based on initial model run results, chose 0.001 percent for the input to the model.

B.3.4 Prior Distributions for Parameters

As discussed elsewhere, parameters in Bayesian models are random variables characterized by probability distributions. Initially, the researcher chooses a probability distribution for each estimated model parameter based on previously available information. These are referred to prior distributions or, simply, "priors." In the case of multi-parameter models, a joint prior distribution captures expected correlations among these parameters.

Once prior distributions are defined, the method of Bayesian inference uses data to update them. The result is a joint "posterior" probability distribution for all the model parameters, one that combines information from the prior distribution and the data to describe the likely range of true parameter values and relationships among these values.

This Bayesian framework allows for expert opinion, independent of the data, to impact parameter estimates by way of the prior distributions. It is also possible, however, to choose prior distributions that have little or no influence on results. This latter approach, which is driven almost entirely by the data, was adopted in this modeling effort. Broad prior distributions were chosen to reflect considerable uncertainty about parameter values at the outset of the surveys.

Given this use of these "minimally informative" prior distributions, it is important to emphasize that these priors are not really an "input" to the model in the same sense as the ICR data, the simulated recovery values, and the tuneable model parameters discussed above. Instead, these priors are more accurately thought of as a flexible structure on top of which parameter estimates are built.

B.3.4.1 Prior Distributions for *Cryptosporidium* Modeling

The next two sections document the prior distributions used in *Cryptosporidium* modeling. Note that these parameter values are on the log-scale for concentration. So, for example, the prior distribution for β_0, the overall mean concentration in the model, is centered at zero, or $10^0 = 1$ oocyst/100L in terms of actual concentration.

Another potentially confusing concept is the specification of variances or "spread" parameters for the prior distributions in the model. There are two key points to keep in mind. First, because it makes the Bayesian math easier, these spread parameters are defined in terms of the distribution's *precision*, which is the inverse of the variance:

precision = $1/\sigma^2$

The most intuitive measure of spread, the standard deviation, is related to precision as follows:

standard deviation = σ = $1/\text{precision}^{1/2}$

Note that this inverse proportion means that larger precision values correspond to smaller standard deviations, and vice versa.

Second, in the model the four precision parameters—prec_1 through prec_4—are "meta-parameters" that are not describing any real-world variation in concentration. Instead, they characterize uncertainty in estimated parameter values. For example, prec_4, the spread parameter for the prior distribution of β_0, the overall mean concentration, does *not* characterize the spread of plant mean concentrations around β_0 but, instead, uncertainty in our estimate of β_0's true value.

Following each prior definition listed below, the range in parentheses captures, roughly, the 1st and 99th percentiles of the distribution. The corresponding 1st and 99th percentiles of the standard deviation computed from the precision as shown above are provided in the brackets. These ranges show that these prior distributions, for the most part, define a very broad range of possible parameter values, and that the prior probability is roughly equal across these ranges. Because there is so little information in these prior distributions, the resulting parameter estimates are driven almost entirely by the data.

This model for log concentration is defined in section B.2.3. Prior distributions for the model parameters are defined below (see previous section for explanation):

β_0 ~ Normal(μ = 0, precision = prec_4), ($10^{-3.2}$ to $10^{+3.2}$), Overall mean concentration
β_1 ~ Normal(μ = 0, precision = prec_4), ($10^{-3.2}$ to $10^{+3.2}$), Slope for standardized turbidity
$\beta_{2:6}$ ~ Normal(μ = 0, precision = prec_4), ($10^{-3.2}$ to $10^{+3.2}$), MSW class effects
ε_i ~ Normal(μ = 0, precision = prec_1), ($10^{-1.2}$ to $10^{+1.2}$), for i = 1 to number of plants
ε_j ~ Normal(μ = 0, precision = prec_2), ($10^{-1.2}$ to $10^{+1.2}$), for j = 1 to number of months
ε_{ij} ~ Normal(μ = 0, precision = prec_3), ($10^{-5.2}$ to $10^{+5.2}$), for all i,j

prec_1 ~ Gamma(α = 2, τ = 0.2), (0.7, 33); [0.17, 1.2]
prec_2 ~ Gamma(α = 2, τ = 0.2), (0.7, 33); [0.17, 1.2]
prec_3 ~ Gamma(α = 2, τ = 4), (0.04, 1.6); [0.79, 5.0]
prec_4 ~ Gamma(α = 2, τ = 2), (0.08, 3.3); [0.55, 3.5]

B.3.4.2 Comparison of Prior and Posterior Distributions

Exhibit B.3 provides a comparison of prior distributions used in the modeling with the resulting posterior distributions.

In this comparison, we expect to see an extreme contrast, with the posterior being much narrower than the original prior. When this happens, it suggests that model parameter estimates are insensitive to our choice of prior distribution (that is, they are based largely on the data). When this is not the case—when the posterior resembles the prior—we question whether our choice of prior has had undue influence on resulting parameter estimates. In Exhibit B.3, some prior distributions are so broad, relative to the resulting posterior distributions, that we see only a small portion of the prior distribution in the plot (for example the flat line at the bottom of the posterior distribution for e_i).

Exhibit B.3 Comparison of Prior Distributions with Resulting Posterior Distributions

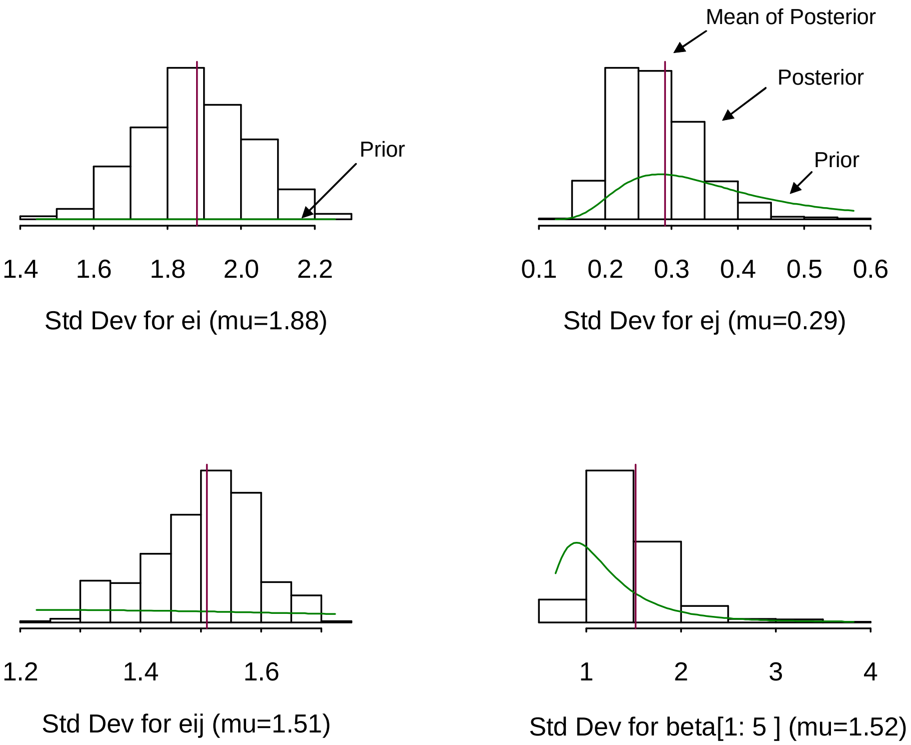
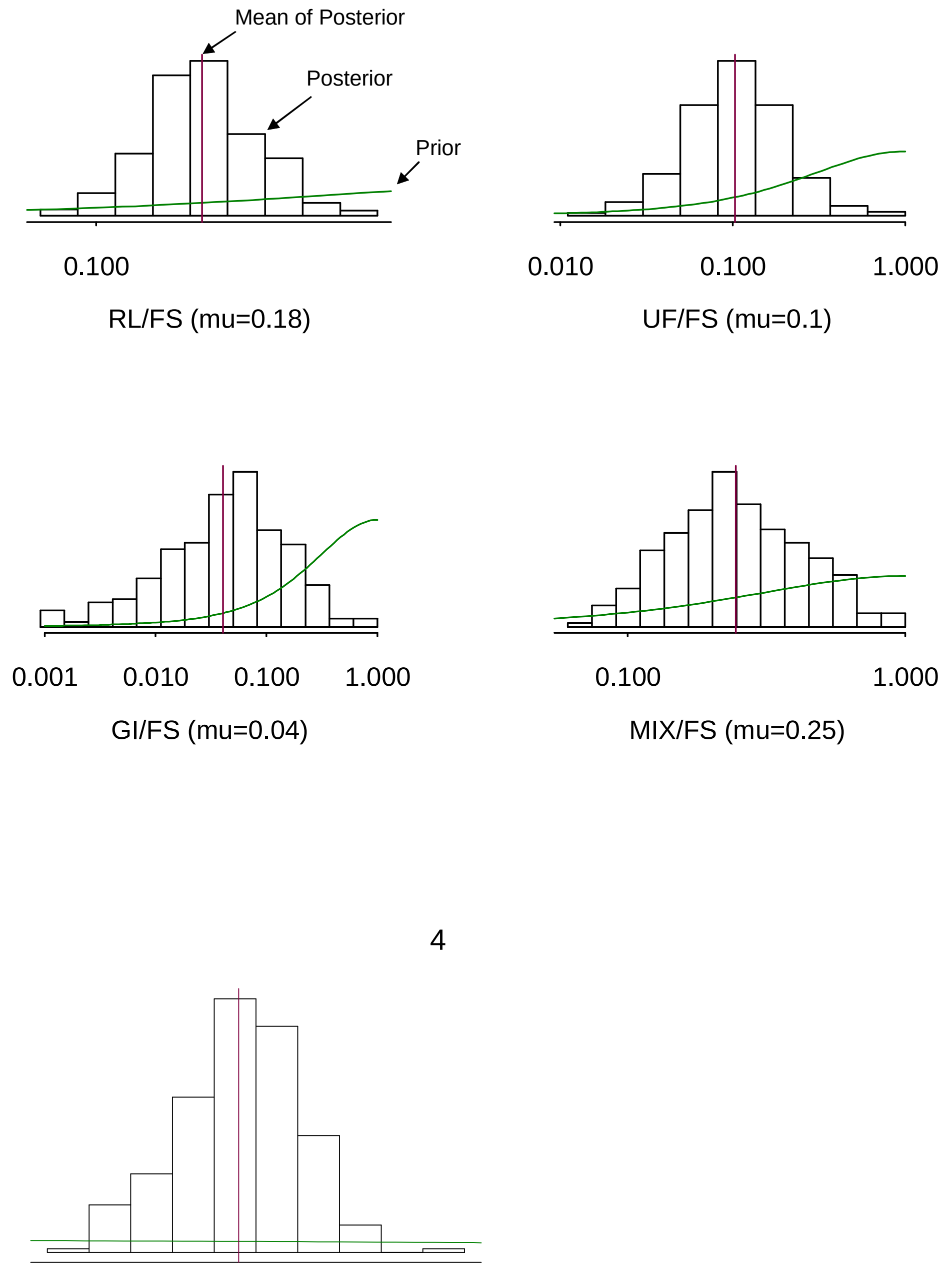


Exhibit B.3 (Continued)
Comparison of Prior Distributions with Resulting Posterior Distributions



RL/FS (mu=0.18)

UF/FS (mu=0.1)

GI/FS (mu=0.04)

MIX/FS (mu=0.25)

Concentration Change with 10-fold Turbidity Increase (mu=1.79)

B.4 Model Fitting and Outputs

The hierarchical Bayesian model was fit to ICR and ICR Supplemental Survey data using an iterative technique known as Markov Chain Monte Carlo (MCMC). This computationally intensive method was carried out using WinBUGS, a software platform developed jointly by the UK Medical Research Council, Biostatistics Unit and the Imperial College School of Medicine at St. Mary's, London. WinBUGS is documented at: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml.

At each step in the bayesian model using MCMC, a single value is drawn from the current distribution for each parameter, with each draw, in turn, conditional on the current distributions for all the other parameters. So the result of a single step is a complete set of parameter values, and, in the limit, these sets of sampled parameter values converge, in distribution, to the target posterior distribution.

There is typically an initial "burn-in" period in which the model fitting takes place, and the sampled values from these iterations are eventually discarded. In our *Cryptosporidium* modeling, this burn-in period was always 400 iterations.

After the burn-in period, when the algorithm has converged to a stable distribution, sampled values are retained, and these empirical distributions (parameter values sampled from the posterior distribution) are used to derive parameter estimates. In *Cryptosporidium* modeling, 5,000 iterations were run following the burn-in period and, from these values, every tenth set was saved (to avoid auto-correlation between values) for a total sample size of 500 per parameter.

B.4.1 Plant-Mean Distributions

Exhibits 4.10, 4.11, and 4.14 summarize modeled *Cryptosporidium* plant-means from each of the three primary occurrence data sets. To obtain these estimates, the average plant-mean concentration is computed at each sampled model iteration, from the set of 12 concentrations drawn from the bayesian model using MCMC for each plant in that iteration. The result is 500 distributions of the average plant-mean concentrations. Observed values from the surveys were used to obtain the simulated sets of 12 samples for each plant that were then used to calculate the means for the plants. The estimates of the means of each plant were then used to compute the parameters of the lognormal distribution of means characterizing plant to plant variability. Uncertainty in the true distribution is reflected by the set of 500 such distributions that are generated. In the exhibits, then, each table row gives the overall mean, median, and 90th percentile from these 500 distributions.

B.4.2 Plant-Mean Distribution Curves

Chapter 4 and Appendix E present cumulative distribution curves for plant-mean concentrations (e.g., Exhibit 4.9). These are obtained from the model, again, through the bayesian model using MCMC sampling algorithm. At each iteration, a mean concentration is computed as described above in B.4.1 for each plant based on the current sample-set of parameter estimates. This collection of sampled plant means is then compared to 41 reference concentrations, one at a time, and the proportion of plant-means falling below each of these reference concentration is computed. The result is a 41-point cumulative distribution curve from each bayesian model using MCMC iteration, or a total of 500 such curves.

Each of the cumulative distribution plots in Chapter 4 and Appendix E summarizes one such set of 500 cumulative distribution curves. The center curve in each plot connects the median (middle) value, across the 500 sampled values, at each of the 41 reference concentrations. In the same way, the dotted-line curves connect the 5th and 95th percentiles.

B.5 Model Evaluations

In each analysis, a single data set (ICR, ICRSS Large Systems, or ICRSS Medium Systems) was used to estimate occurrence of either Cryptosporidium or Giardia in drinking water sources. Each analysis produced a large sample of occurrence distributions. Each individual occurrence distribution provides one plausible picture of the variability of average concentration among the nation's drinking water sources. A large "sample" of such distributions reveals uncertainty, due to limited data in the "true" distribution of variability. The following subsections discuss mixing and autocorrelation, internal and external model checks, and checks for bias in annual estimates due to seasonality.

B.5.1 Mixing and Autocorrelation

Because the bayesian model using MCMC fitting is an iterative process, sampled parameter values from nearby iterations are often correlated. When present, this auto-correlation can result in parameter distributions that systematically under-estimate the variance of the target posterior distribution. To avoid this problem, it is standard practice in the bayesian model using MCMC sampling to skip iterations between samples. In the modeling documented here, samples were always drawn from every tenth iteration, since lag plots consistently showed little evidence of auto-correlation at this spacing.

B.5.2 Internal Model Check

Exhibits 4.9, 4.12, and 4.13 summarize the fit of modeled plant-mean distributions to observed sample distributions. In each exhibit, one for each of the three primary data sets, the dashed line shows the distribution of observed plant-mean concentrations. As discussed in Section B.4.2, the solid line and dotted lines, together, summarize a collection of 500 modeled occurrence distributions.

In the lower half of each distribution, the effect of limited sample volumes is clear. Modeling predicts smooth distributions through these very low concentrations, while the observed distribution curve is constrained by "detection limits", and never drops below the overall proportion of zero-count locations. In the upper half of each distribution, though, the observed data curves generally fall within the 90 percent modeled limits suggesting a good fit, model to data.

B.5.3 External Model Check

To investigate the predictive value of the *Cryptosporidium* modeling, the following external model check was carried out:

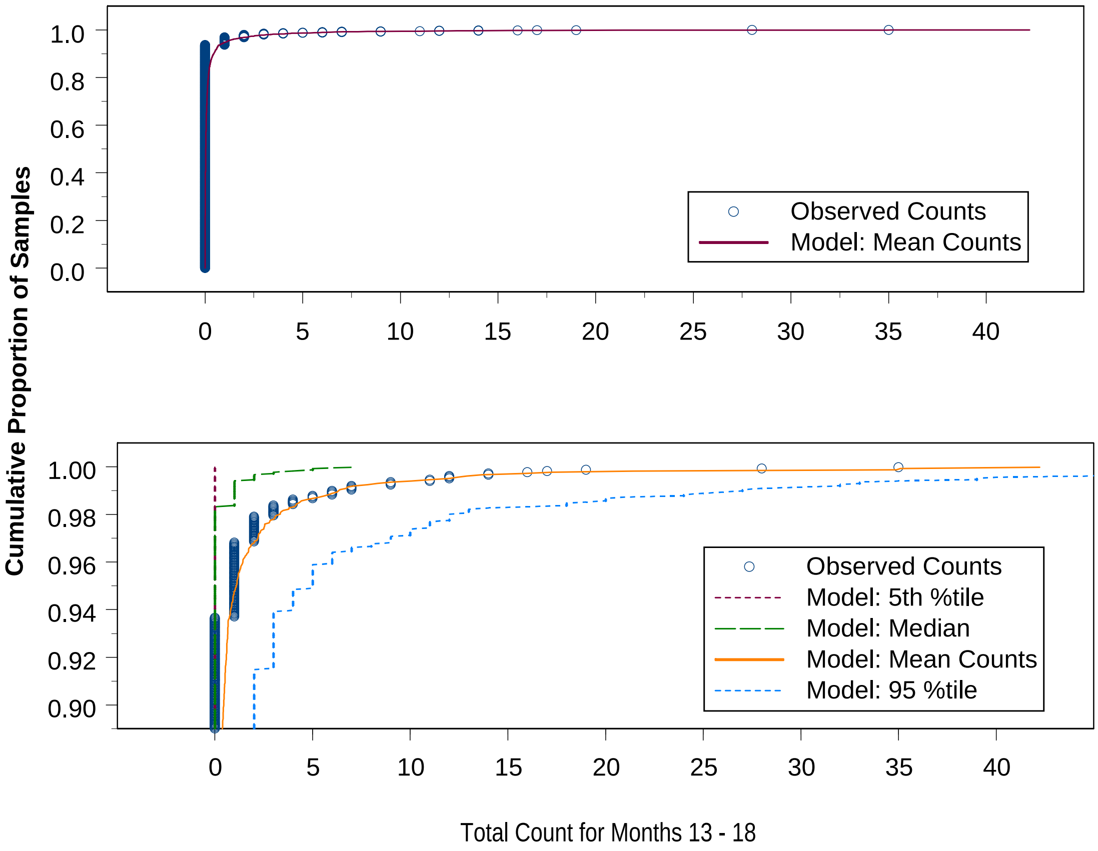
1) Fit the model (Section B.3) to the first 12 months of ICR *Cryptosporidium* data only.

2) Use the fitted model from Step 1 along with the input values from months 13 to 18 (sample volume, turbidity, etc) to predict oocyst counts for months 13 to 18.

3) Compare the predicted counts for months 13 to 18 to the actual, observed counts for these months.

Results are summarized in Exhibit B.4 in the form of cumulative count distributions. On both plots, the circles represent the actual sample counts over the last 6 months of the ICR. The various lines capture different statistics for the modeled counts. Agreement is good.

Exhibit B.4 Results from External Model Check



B.5.4 Check for Bias in Annual Estimates Due to Seasonality

Microbial concentrations are thought to be related to the frequency and intensity of rainfall, especially for water systems that are fed by flowing rivers and streams. In attempting to estimate annual occurrence from an 18-month survey, such effects introduce the potential for seasonal bias. This is because, for each location in the survey, we capture one complete annual cycle plus one half-year block. Unless the half-year block fairly represents the typical full-year cycle at a given location (is not disproportionately from any season), we run the risk of over or under-estimating annual occurrence at this location. In going from individual location estimates to a national distribution of plant means, such errors would have to "cancel" each other perfectly to avoid bias.

In Section 4.6, we present some evidence of seasonal trends in *Cryptosporidium* occurrence. As a result, we are not able to rule out the possibility of such a bias, due to seasonal effects, in our estimates of annual average *Cryptosporidium* occurrence rates from the ICR. Note that this potential problem is only relevant to the 18-month ICR Survey. The ICR Supplemental Survey was carried out over a 12-month period, covering one annual cycle.

This section summarizes attempts to measure how big such a bias might be and whether it could have a significant impact at the next level, where *Cryptosporidium* occurrence models serve as input to the LT2ESWTR Economic Analysis. The basic approach is to construct alternative, unbiased estimates of the national plant-mean distribution based on 12-month intervals. Because they are based on less data, these alternative estimates are in some ways inferior to our primary 18-month estimate, but they are free from potential bias due to seasonal effects. We then compare these alternate distributions, based on 12 consecutive months, to our primary distribution, based on 18 consecutive months, to assess the likely magnitude of any such bias.

Within the 18-month ICR monitoring period, there are seven overlapping 12-month intervals: July 1997 to June 1998, August 1997 to July 1998, … , and January 1998 to December 1998. Estimates of occurrence based on any one of these intervals will be free of any bias from seasonal effects, since each captures one complete cycle of seasons.

There are two ways to obtain these 12-month plant mean distributions. In the first approach, model parameters are estimated using all 18 months of data. Since the model includes a set of parameters that measures each monthly effect, it is possible to construct plant-mean estimates by month from these 18-month parameter estimates, and then group these monthly means into consecutive 12-month collections. The second method is to simply subset the data into 12- month periods and model each period separately. Since there are pros and cons to each approach, the model check was carried out both ways.

In both cases, the comparison of 18-month and 12-month plant-mean curves will be confounded, to some extent, by the differences in number of months sampled (n=12 means will vary more than n=18 means, all other things being equal). Also, the second approach might show slightly more spread in estimated plant means than the first due to smaller-data-set parameter estimates.

Exhibit B.5 shows the results from the first method, and Exhibit B.5 results from the second. Both show little difference between the 12-month and 18-month distribution curves. Although differences are clearly greater in the second approach (Exhibit B.6), they are small enough to be caused by the sample-size effects discussed above.

Exhibit B.5 Annual CDFs Constructed From Parameter Estimates Based on All Data

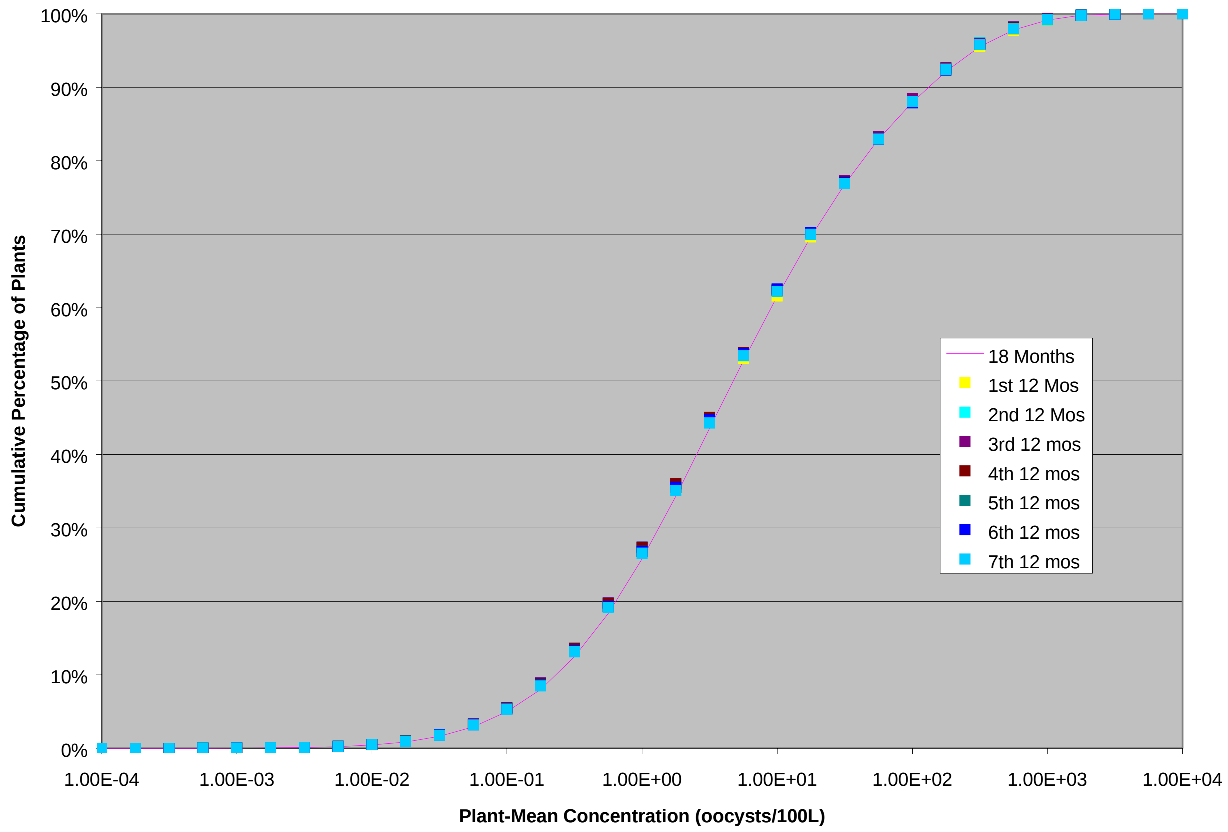
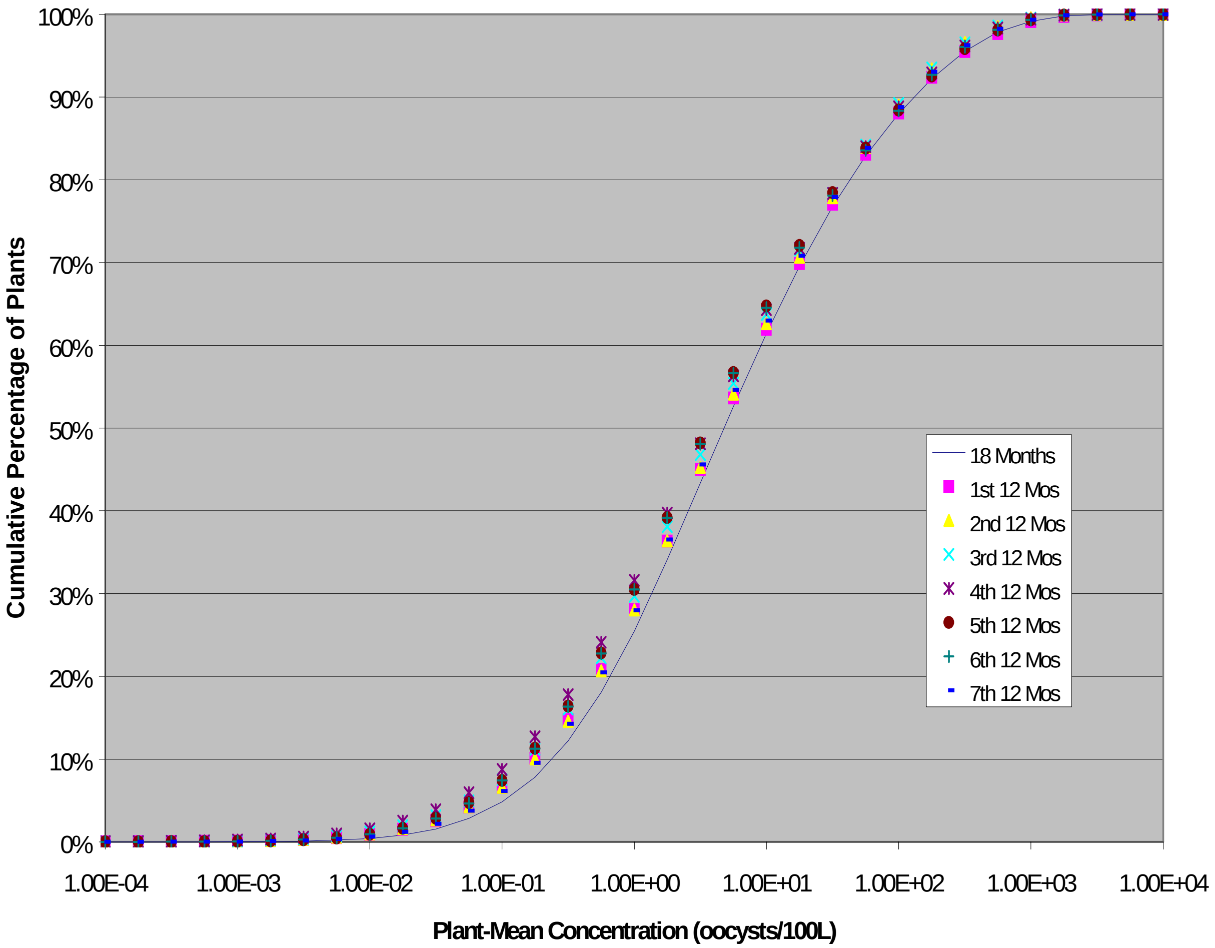


Exhibit B.6 Annual CDFs Based on 12-Month Data Sets



B.6 Reduced-Form Model

The occurrence model described in the preceding sections of this appendix was used to develop the information for filtered systems in Chapters 3 and 4. EPA has also developed and implemented a reduced form of the model (also referred to as the “simplified” model), which was used to provide the information in Chapters 3 and 4 on unfiltered systems. The output of the reduced model was also used as the direct input to the cost and benefit analyses for both filtered and unfiltered systems in the economic analyses of LT2ESWTR regulatory options.

The reduced form of the model was developed because of limitations observed in the national occurrence distributions for unfiltered systems generated by the full form of the model. While those distributions appeared reasonable across most of the range, it appeared that the upper tails were overstating the possible occurrence of average *Cryptosporidium* levels in source water used by unfiltered systems. EPA considered a number of alternatives, such as truncating or modifying the upper tail of the unfiltered system distributions, including some modifications to the model to reflect some particular aspects related to the data available for the unfiltered systems.

Unfiltered systems are locations that, at the time of the ICR survey, met strict source water purity standards that excluded them from the regulatory requirement to filter. In light of this prior knowledge, it makes sense to estimate occurrence independently for this class of system. The ICRSS also included a few unfiltered plants, but too few to model separately, leaving ICR data as the only useful source for unfiltered system occurrence estimates.

The full model described in the previous section includes a large number of parameters: six β terms, a second set of parameters representing every location in the survey, a third representing every sample month, and, finally, an even larger collection of residual terms. The large number of filtered plants in the ICR survey (n=338) provides enough data to comfortably estimate all these independent parameters. However, the data from the much smaller sample of unfiltered plants (n=12) is too sparse. The number of modeled parameters begins to approach the number of independent data points, diminishing the usefulness of parameters as representations of more general patterns.

Moreover, measured turbidity values for these unfiltered locations are all very low, with little variation among them. This makes sense in light of the regulatory requirements for avoiding filtration. On the standardized scale, the average plant-mean turbidity for ICR unfiltered plants was -1.5, versus 0.08 for filtered plants.

For the economic analysis of the LT2ESWTR, EPA used the reduced form model to predict both filtered systems' and unfiltered systems' occurrence distributions. While the simple model was initially developed for unfiltered systems, it was also able to produce the data needed as input to the risk assessment model. Note, the risk assessment model for the LT2ESWTR uses 1,000 log-normal distributions of plant-mean to reflect both variability and uncertainty in *Cryptosporidium* national occurrence (plant to plant variability in each distribution, uncertainty from the set of 1,000 of these distributions). Each of the 1,000 distributions of plant means represents a plausible national distribution of plant means based on underlying data. EPA compared the occurrence estimates of the full model and simple model when considering which to use in the economic analysis, and determined there was no significant difference between the two.

The following provides additional detail about the reduced-form model and the comparison between the full model and reduced-form model.

B.6.1 Expected Counts in Reduced-Form Model

At this level, the reduced-form model is the same as the full model. The observed counts, $Y_{i,j}$, are modeled as Poisson random variables and the expected counts are built from concentration, volume, and recovery:

$$Y_{i,j} \sim \text{Poisson}(\lambda_{i,j})$$

$$\lambda_{i,j} = \text{Concentration}_{i,j} \times \text{Volume}_{i,j} \times \text{Recovery}_{i,j}$$

The only difference here is the lack of an assumed false positive contribution (FP) to the count means. In earlier work with the full model, the impact of this false positive term was negligible over the range of likely values, so it was dropped from this simpler model.

B.6.2 Reduced-Form Model Estimates of the Distribution of Plant-Mean Concentrations

In the full model, estimated concentrations are broken down into a number of underlying effects and parameters are estimated to model the impact of each general effect on concentration. In the reduced-form model, the focus shifts to modeling the distribution of estimated concentrations, both within a particular location (over time) and from location-to-location.

The reduced model is similar to the full model described in B.2.3:

$$\text{Concentration}_{i,j} = \exp(\beta_{Filt} + \beta_1 * \text{ Unfiltered?}_{i,j} + \varepsilon_i + \varepsilon_{ij}),$$

Where:

- β_{Filt} is an intercept term, the median occurrence level among filtered plants
- β_1 is a fixed effect for plants that do not filter. A negative value for this parameter would predict lower median occurrence in the source waters of plants that filter.
- Unfiltered?$_{i,j}$ = 1 if plant i does not use filtration during month j and 0, otherwise.
- ε_i = random effect for location. This allows different source waters to have different occurrence levels
- ε_{ij} = residual term that embodies other variation

The reduced model is simpler because it predicts different occurrence levels for only two types of water (filtered and unfiltered) and includes no effects for months. There are only two precision terms, prec_1 describing normally distributed ε_i and prec_2 describing normally distributed ε_{ij}.

Thus, each filtered plant will have geometric mean defined by β_{Filt} and the plant-specific effect, ε_i. Concentrations at filtered plant i would be lognormally distributed so that the natural log concentration is normally distributed with mean $\beta_{Filt} + \varepsilon_i$ and variance $1/\text{prec}_2$.

$$\ln(C_{i,j}) \sim N(\beta_{Filt} + \varepsilon_i, \text{prec}_2)$$

The random effects ε_i describe variability from location-to-location among plants that filter and also among unfiltered plants. These are normally distributed about zero with variance $1/\text{prec}_1$.

$$\varepsilon_i \sim N(0, \text{prec}_1)$$

Within a plant (either filtered or unfiltered), random effects $\varepsilon_{i,j}$ describe how concentration varies over time. Within a plant, these effects are normally distributed with variance $1/\text{prec}_2$.

$$\varepsilon_{i,j} \sim N(0, \text{prec}_2)$$

For unfiltered plants, the model equations are nearly identical. Unfiltered plants will have medians defined by $\beta_{Unfilt} = \beta_{Filt} + \beta_1$. In the discussion that follows, priors, likelihoods, and estimates for unfiltered systems are expressed in terms of β_{Unfilt} rather than $\beta_{Filt} + \beta_1$.

B.6.3 Reduced-Form Model Prior Distributions

In the reduced-form model, it is necessary to use prior distributions for the fixed effects (β_{Filt} and β_{Unfilt}) and precisions ($prec_1$ and $prec_2$). As in the full model, these prior distributions were selected to be as broad as possible, reflecting genuine uncertainty about these true values and allowing the data to drive their posterior distributions.

Prior distributions for the model parameters are defined below:

β_{Unfilt} ~ N(0,σ^2 = 10,000)
β_{Filt} ~ N(0, σ^2 = 10,000)
prec1 ~ Gamma(α= 2, τ= 0.2), which has 98% probability mass in [0.7, 33]
prec2 ~ Gamma(α= 2, τ = 2), which has 98% probability mass in [0.08, 3,3]

In the initial modeling work, separate precision terms were assigned to filtered and unfiltered plants. Although that setup was theoretically reasonable, the small number of unfiltered plants failed to support estimating their precision parameters. There was insufficient evidence to reject the notion that unfiltered and filtered plants have common precision parameters (i.e., the hypotheses that the two kinds of systems have equal between-plant and within-plant variances could not be rejected). As a result, parameter $prec_1$ describes between-plant variability, while parameter $prec_2$ describes within-plant variability for both filtered and unfiltered plants.

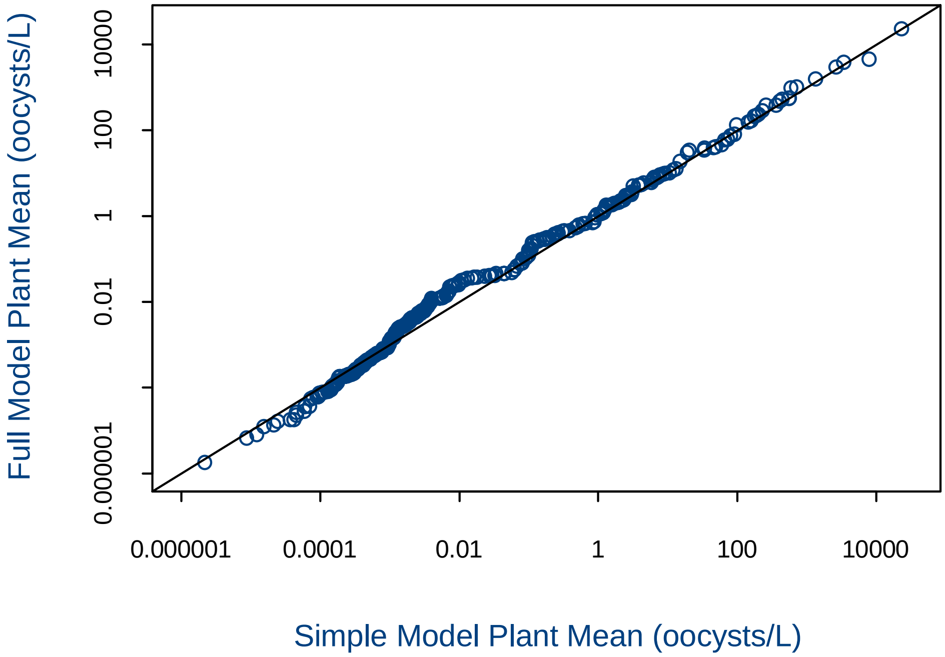
Accordingly, the model uses all of the data (both filtered and unfiltered) to estimate these two precision parameters. The overall group medians (β_{Filt} and β_{Unfilt}), however, are estimated independently.

B.6.4 Comparison of Full Model to Reduced-Form Model

To ensure that the reduced-form model did not differ from the full model in a manner that would affect the estimated plant-mean distributions used in subsequent economic analyses, EPA conducted a comparative analysis. To capture uncertainty and variability of the occurrence estimates, the risk model uses 1,000 plant-mean occurrence distributions. These distributions comprise the individual plant-means (350 plant-means for the ICR data set and 40 for each the ICRSSL and ICRSSM data sets). Exhibit B.7 shows the individual plant-means predicted by the full model versus the simple model and indicates how closely the two models correlate, with respect to plant-mean estimates.

The plant-means fall on the line where the full model and simple model predicted the same values. There is a small difference around the 0.01 oocyst/L level, where the simple model predicts slightly lower concentrations than the full model does.

Exhibit B.7 Comparison of Full and Simple Occurrence Models



Appendix C. Boxplots of Observed ICR Data

Observed data are graphically presented using the following distributions:

- Boxplots of monthly distributions (18 months of data)

- Boxplots of annual cumulative distributions (7 running annuals from the 6 quarters of data)

- Annual cumulative distributions (7 running annuals from the 6 quarters of data)

Exhibit C.1 presents an example of the boxplot diagrams used to present the data in Appendix C. The boxplot identifies the mean, median, minimum point, maximum point, and 10th, 25th, 75th, and 90th percentiles. The data points located below the 5th percentile and above the 95th percentile are plotted individually.

At the bottom of each boxplot diagram for the Information Collection Rule (ICR) data, the number of samples taken (N), the number of non-detects (N-Dct), and the standard deviation (Std) are given for each month or year of sampling. If no standard deviation is shown, this means the standard deviation is too large to display (>999). This happens only for coliform bacteria. ICR Sampling began in July 1997 (noted as JL1 in the monthly boxplots) and ended in December 1998 (DC2). The annual boxplots contain all the monthly data for a given 12-month period (e.g., July 1997-June 1998 (J-J), August 1997-July 1998 (A-J)).

Exhibit C.1 Boxplot Diagram

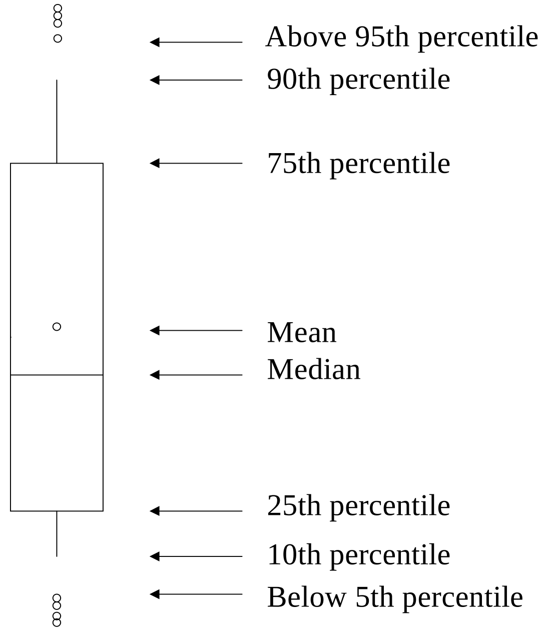
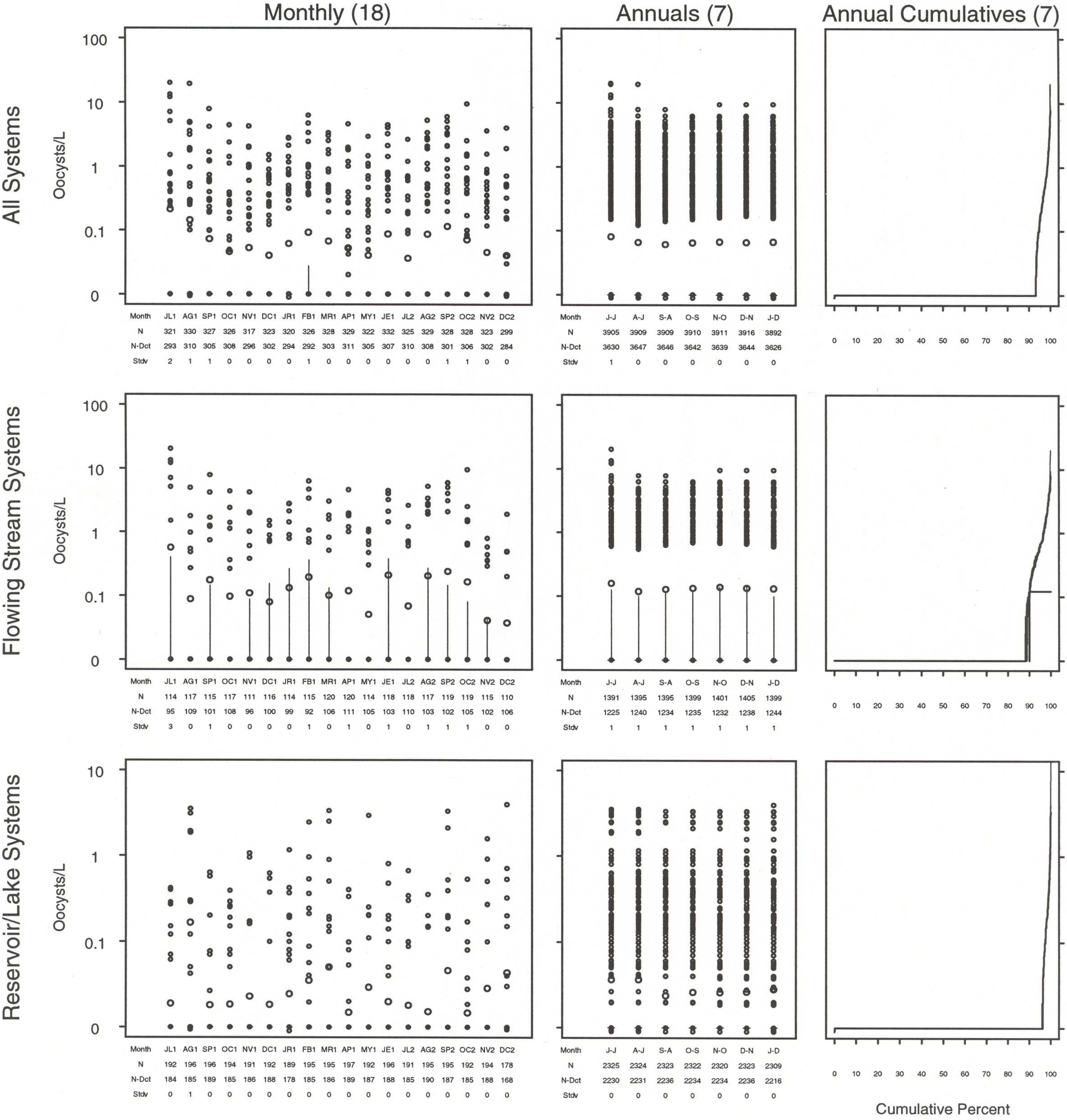


Exhibit C.2 Exhibit List

Exhibit	Pathogen
C-3	*Cryptosporidium* Total
C-4	*Cryptosporidium* Non-Empty
C-5	*Cryptosporidium* - With Internal Structure
C-6	*Giardia* Total
C-7	*Giardia* - Non-Empty
C-8	*Giardia* - Internal Structure
C-9	*Giardia* - Greater than One Internal Structure
C-10	Viruses
C-11	Total Coliform
C-12	Fecal Coliforms
C-13	*E. Coli*

For each protozoan and for coliform bacteria, data are separated by source water type and by filtration status. There was insufficient data to generate boxplots for viruses in unfiltered systems.

Exhibit C-3:
Concentration of Total Cryptosporidium Oocysts in Plant Influent,
ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems, Monthly

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	293	310	305	308	296	302	294	292	303	311	305	307	310	308	301	306	302	284
Stdv	2	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0

All Systems, Annuals

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3905	3909	3909	3910	3911	3916	3892
N-Dct	3630	3647	3646	3642	3639	3644	3626
Stdv	1	0	0	0	0	0	0

Flowing Stream Systems, Monthly

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	95	109	101	108	96	100	99	92	106	111	105	103	110	103	102	105	102	106
Stdv	3	0	1	0	0	0	0	1	0	1	0	1	0	1	1	1	0	0

Flowing Stream Systems, Annuals

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1225	1240	1234	1235	1232	1238	1244
Stdv	1	1	1	1	1	1	1

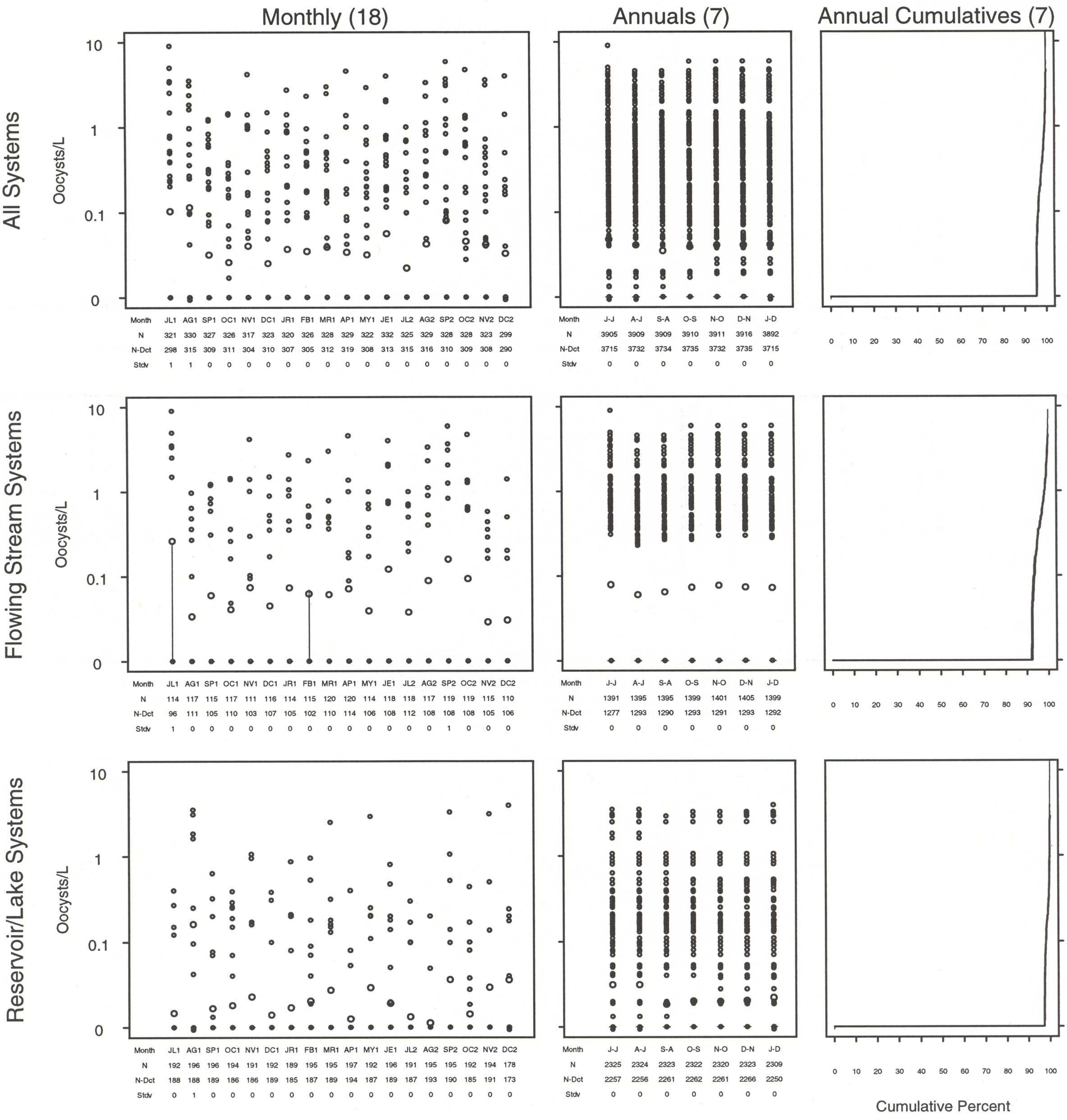
Reservoir/Lake Systems, Monthly

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	184	185	189	185	186	188	178	185	186	189	187	188	185	190	187	185	188	168
Stdv	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Reservoir/Lake Systems, Annuals

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2230	2231	2236	2234	2234	2236	2216
Stdv	0	0	0	0	0	0	0

Exhibit C-4:
Concentration of Non-Empty Cryptosporidium Oocysts in Plant Influent,
ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	298	315	309	311	304	310	307	305	312	319	308	313	315	316	310	309	308	290
Stdv	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3905	3909	3909	3910	3911	3916	3892
N-Dct	3715	3732	3734	3735	3732	3735	3715
Stdv	0	0	0	0	0	0	0

Flowing Stream Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	96	111	105	110	103	107	105	102	110	114	106	108	112	108	108	108	105	106
Stdv	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

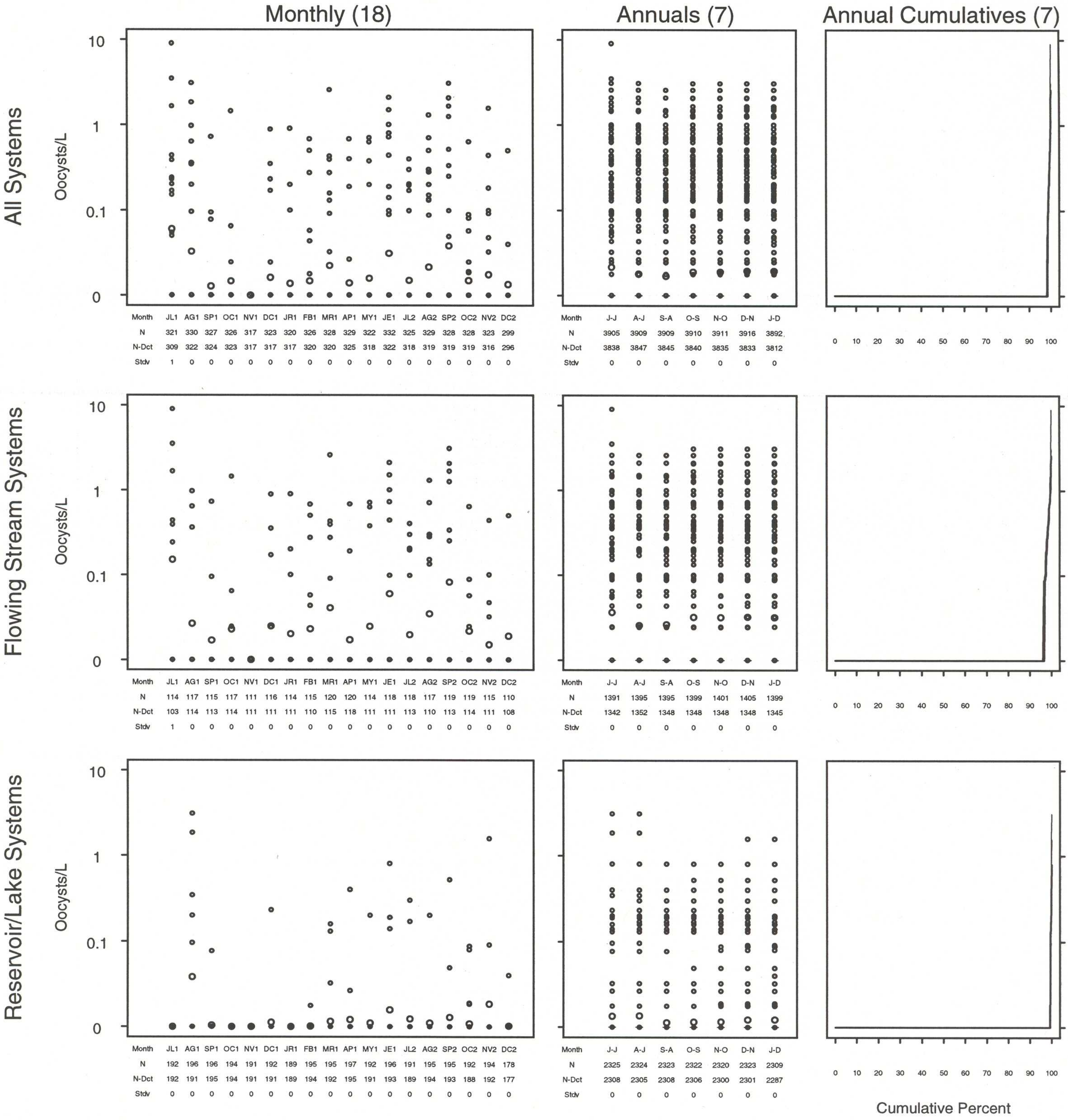
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1277	1293	1290	1293	1291	1293	1292
Stdv	0	0	0	0	0	0	0

Reservoir/Lake Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	188	188	189	186	186	189	185	187	189	194	187	189	187	193	190	185	191	173
Stdv	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2257	2256	2261	2262	2261	2266	2250
Stdv	0	0	0	0	0	0	0

Exhibit C-5:
Concentration of Cryptosporidium Oocysts with Internal Structures in Plant Influent
ICR Plant-Month Data for 18 Months by Surface Water Category
(All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	309	322	324	323	317	317	317	320	320	325	318	322	318	319	319	319	316	296
Stdv	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All Systems – Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3905	3909	3909	3910	3911	3916	3892,
N-Dct	3838	3847	3845	3840	3835	3833	3812
Stdv	0	0	0	0	0	0	0

Flowing Stream Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	103	114	113	114	111	111	111	110	115	118	111	111	113	110	113	114	111	108
Stdv	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Flowing Stream Systems – Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1342	1352	1348	1348	1348	1348	1345
Stdv	0	0	0	0	0	0	0

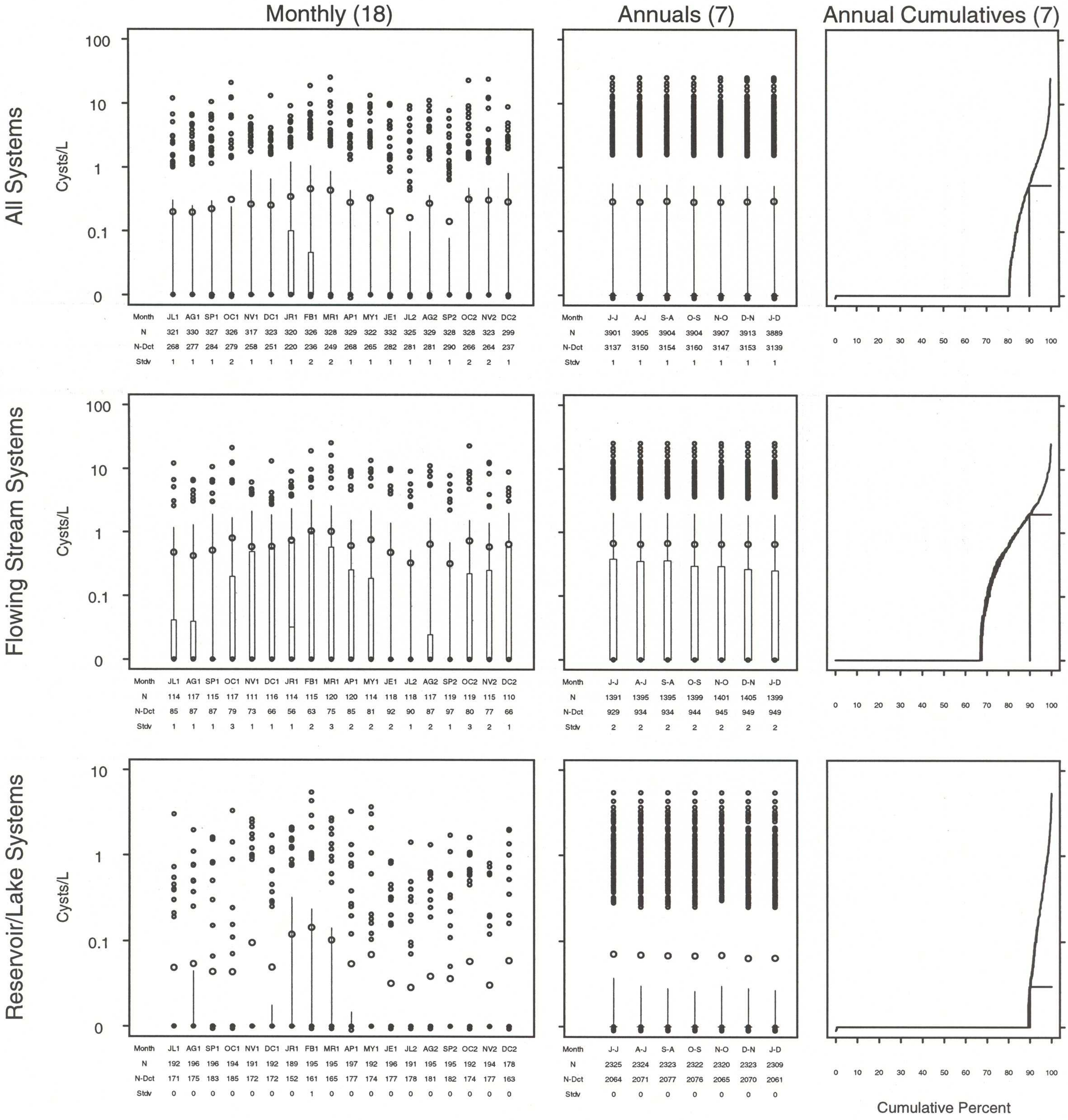
Reservoir/Lake Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	192	191	195	194	191	191	189	194	192	195	191	193	189	194	193	188	192	177
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Reservoir/Lake Systems – Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2308	2305	2308	2306	2300	2301	2287
Stdv	0	0	0	0	0	0	0

Exhibit C-6:
Concentration of Total Giardia Cysts in Plant Influent, ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	268	277	284	279	258	251	220	236	249	268	265	282	281	281	290	266	264	237
Stdv	1	1	1	2	1	1	1	2	2	1	1	1	1	1	1	2	2	1

All Systems – Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3901	3905	3904	3904	3907	3913	3889
N-Dct	3137	3150	3154	3160	3147	3153	3139
Stdv	1	1	1	1	1	1	1

Flowing Stream Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	85	87	87	79	73	66	56	63	75	85	81	92	90	87	97	80	77	66
Stdv	1	1	1	3	1	1	1	2	3	2	2	2	1	2	1	3	2	1

Flowing Stream Systems – Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	929	934	934	944	945	949	949
Stdv	2	2	2	2	2	2	2

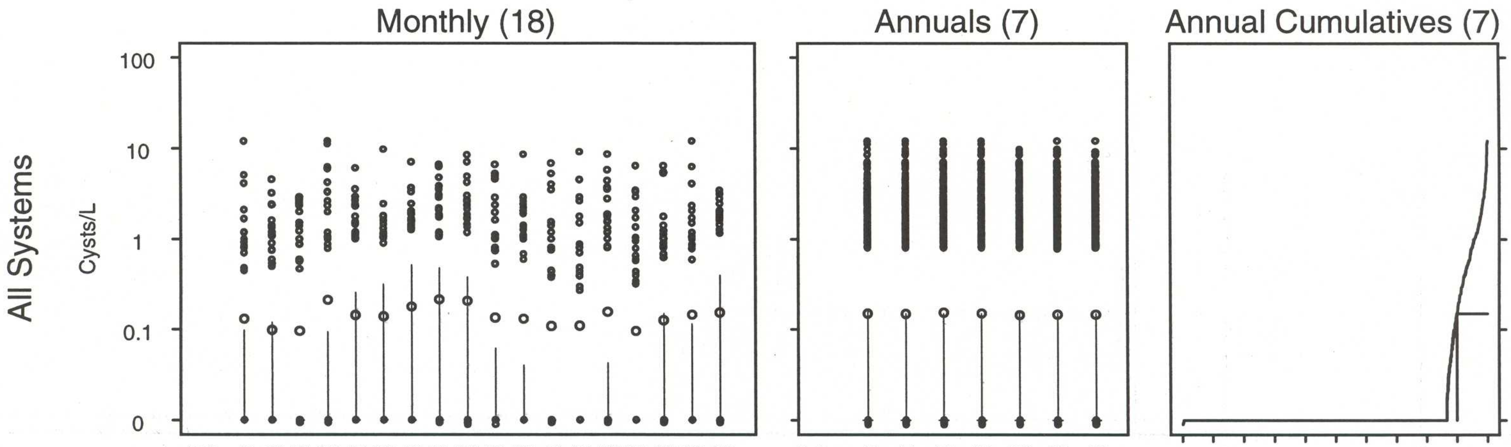
Reservoir/Lake Systems – Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	171	175	183	185	172	172	152	161	165	177	174	177	178	181	182	174	177	163
Stdv	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Reservoir/Lake Systems – Annuals (7)

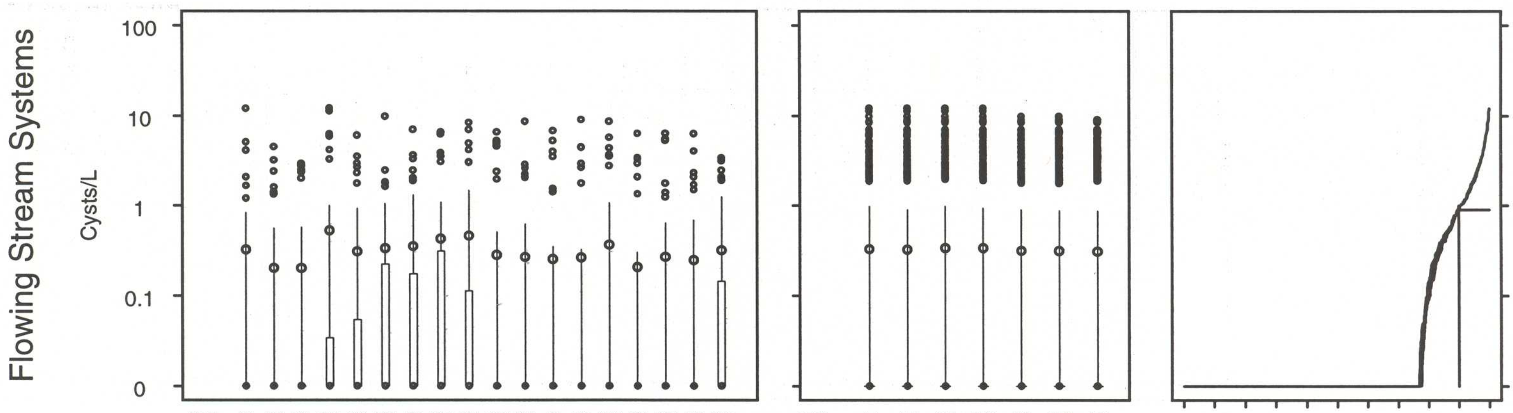
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2064	2071	2077	2076	2065	2070	2061
Stdv	0	0	0	0	0	0	0

Exhibit C-7:
Concentration of Non-Empty Giardia Cysts in Plant Influent,
ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



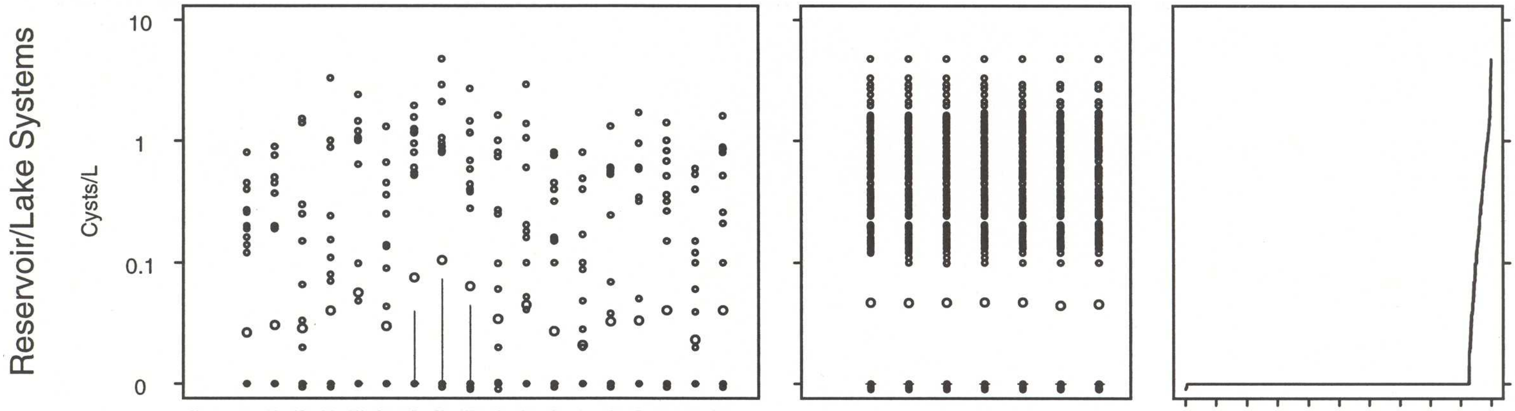
Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	282	294	301	287	279	271	254	257	269	292	284	300	298	294	304	283	282	254
Stdv	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3901	3905	3904	3904	3907	3913	3889
N-Dct	3370	3386	3386	3389	3385	3388	3371
Stdv	1	1	1	1	1	1	1



Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	92	95	99	87	82	80	75	77	85	98	95	102	102	94	104	91	89	78
Stdv	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1

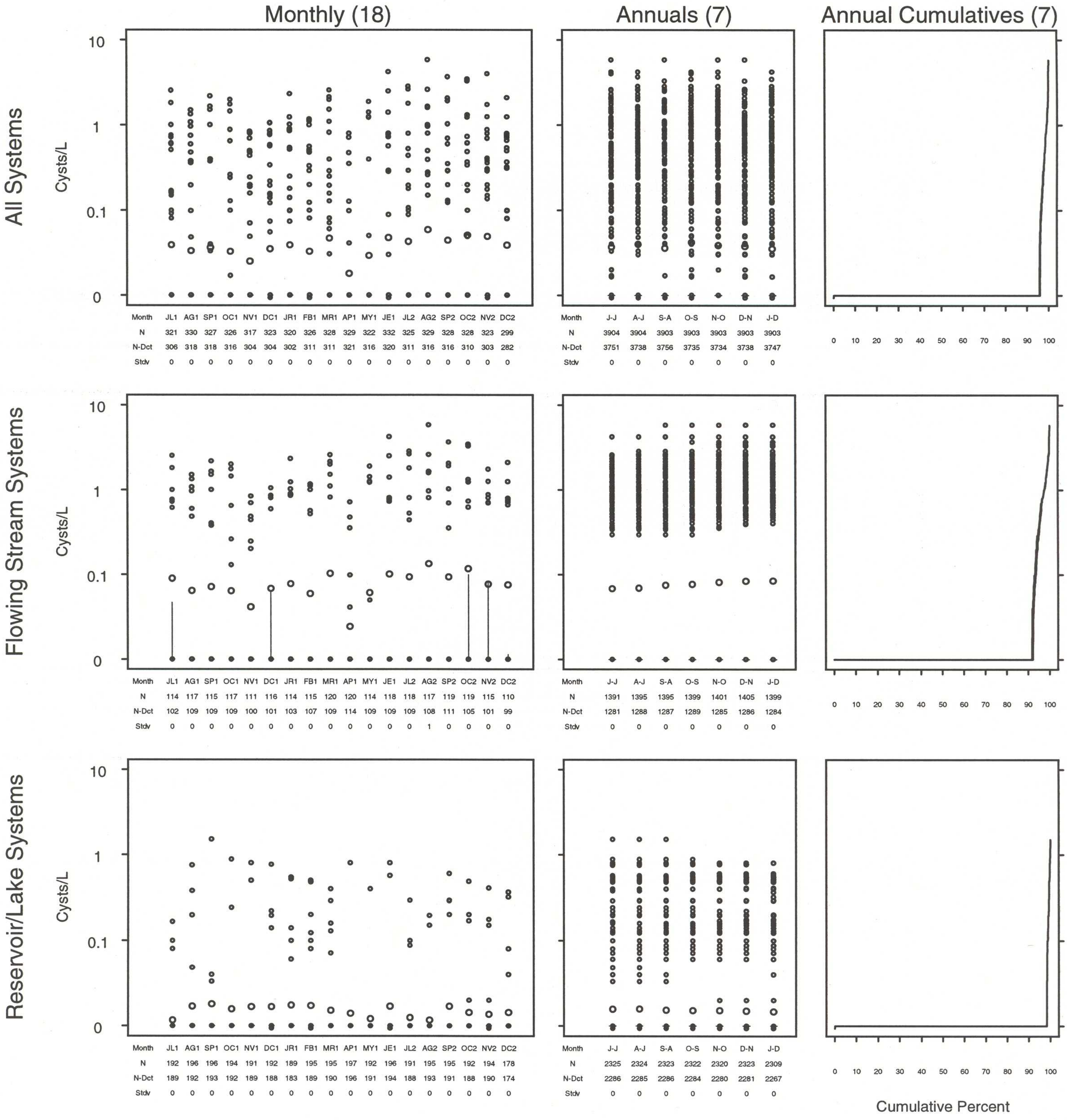
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1067	1077	1076	1081	1085	1092	1090
Stdv	1	1	1	1	1	1	1



Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	177	184	188	185	182	177	166	167	173	187	178	184	182	186	188	179	182	167
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2148	2153	2155	2155	2149	2149	2139
Stdv	0	0	0	0	0	0	0

Exhibit C-8:
Concentration of Giardia Cysts with Internal Structures in Plant Influent,
ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	306	318	318	316	304	304	302	311	311	321	316	320	311	316	316	310	303	282
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3904	3904	3903	3903	3903	3903	3903
N-Dct	3751	3738	3756	3735	3734	3738	3747
Stdv	0	0	0	0	0	0	0

Flowing Stream Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	102	109	109	109	100	101	103	107	109	114	109	109	109	108	111	105	101	99
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

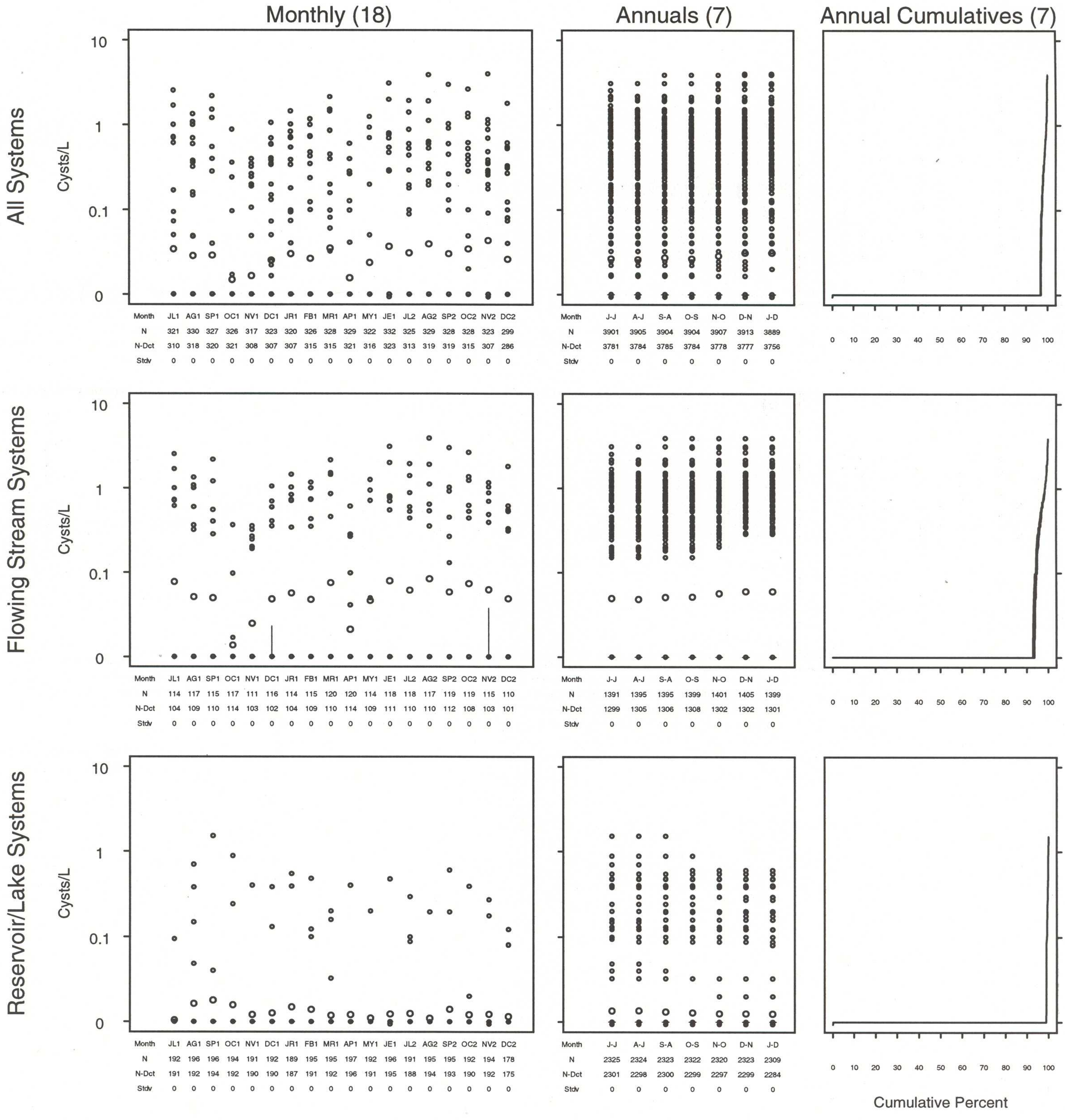
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1281	1288	1287	1289	1285	1286	1284
Stdv	0	0	0	0	0	0	0

Reservoir/Lake Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	189	192	193	192	189	188	183	189	190	196	191	194	188	193	191	188	190	174
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2286	2285	2286	2284	2280	2281	2267
Stdv	0	0	0	0	0	0	0

Exhibit C-9:
Concentration of Giardia Cysts with >1 Internal Structure in Plant Influent, ICR Plant-Month Data for 18 Months By Surface Water Category (All, Flowing Streams, Reservoir Lakes) Monthly & Annual Boxplots, Annual Cumulative Distributions July 1997 - December 1998



All Systems — Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	321	330	327	326	317	323	320	326	328	329	322	332	325	329	328	328	323	299
N-Dct	310	318	320	321	308	307	307	315	315	321	316	323	313	319	319	315	307	286
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

All Systems — Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3901	3905	3904	3904	3907	3913	3889
N-Dct	3781	3784	3785	3784	3778	3777	3756
Stdv	0	0	0	0	0	0	0

Flowing Stream Systems — Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	114	117	115	117	111	116	114	115	120	120	114	118	118	117	119	119	115	110
N-Dct	104	109	110	114	103	102	104	109	110	114	109	111	110	110	112	108	103	101
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Flowing Stream Systems — Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1391	1395	1395	1399	1401	1405	1399
N-Dct	1299	1305	1306	1308	1302	1302	1301
Stdv	0	0	0	0	0	0	0

Reservoir/Lake Systems — Monthly (18)

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	196	196	194	191	192	189	195	195	197	192	196	191	195	195	192	194	178
N-Dct	191	192	194	192	190	190	187	191	192	196	191	195	188	194	193	190	192	175
Stdv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Reservoir/Lake Systems — Annuals (7)

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2325	2324	2323	2322	2320	2323	2309
N-Dct	2301	2298	2300	2299	2297	2299	2284
Stdv	0	0	0	0	0	0	0

Exhibit C-10:
Concentration of Viruses in Plant Influent, ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions,
July 1997 - December 1998

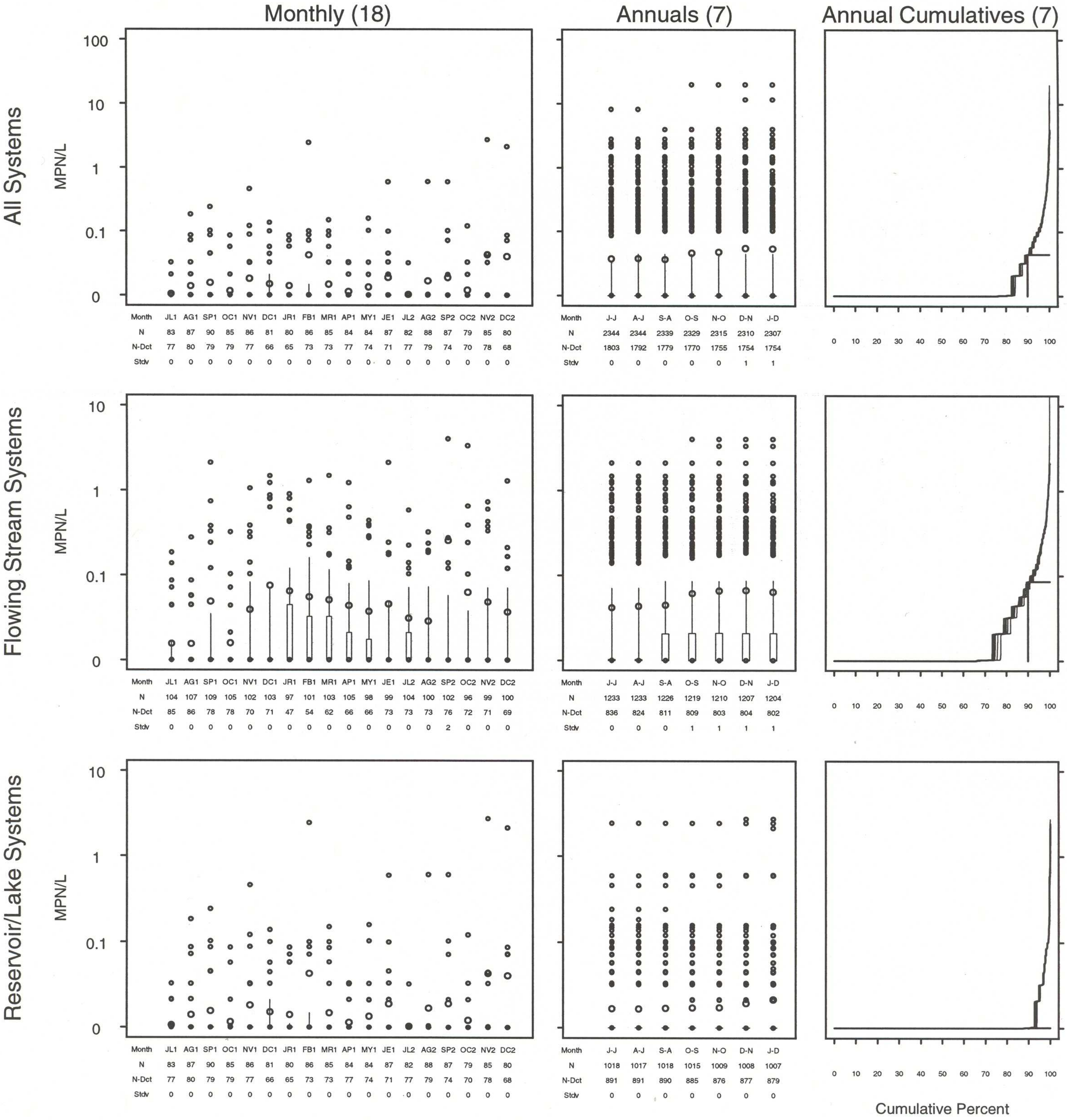
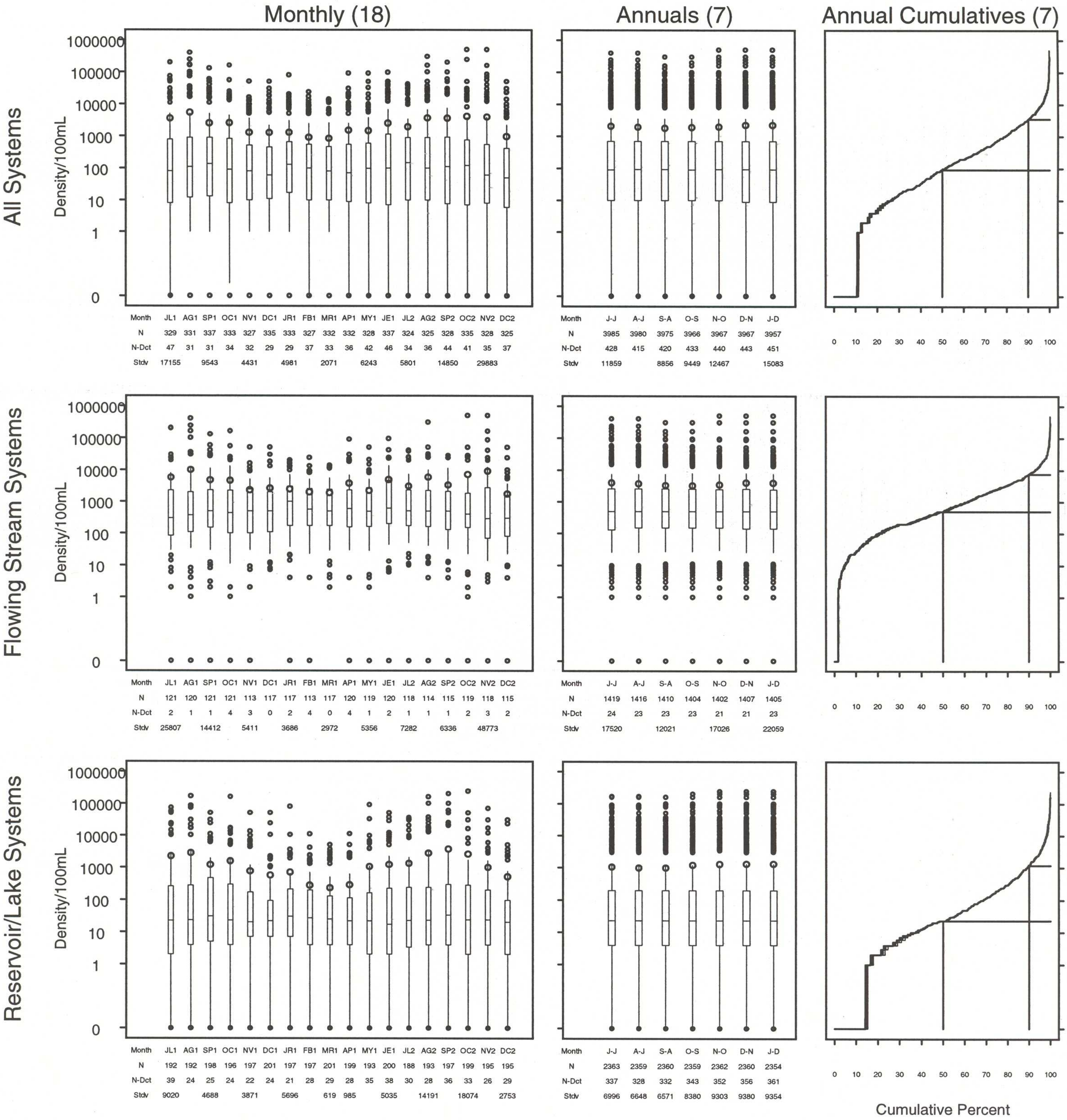


Exhibit C-11:
Concentration of Total Coliform in Plant Influent, ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	329	331	337	333	327	335	333	327	332	332	328	337	324	325	328	335	328	325
N-Dct	47	31	31	34	32	29	29	37	33	36	42	46	34	36	44	41	35	37
Stdv	17155		9543		4431		4981		2071		6243		5801		14850		29883	

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	3985	3980	3975	3966	3967	3967	3957
N-Dct	428	415	420	433	440	443	451
Stdv	11859		8856	9449	12467		15083

Flowing Stream Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	121	120	121	121	113	117	117	113	117	120	119	120	118	114	115	119	118	115
N-Dct	2	1	1	4	3	0	2	4	0	4	1	2	1	1	1	2	3	2
Stdv	25807		14412		5411		3686		2972		5356		7282		6336		48773	

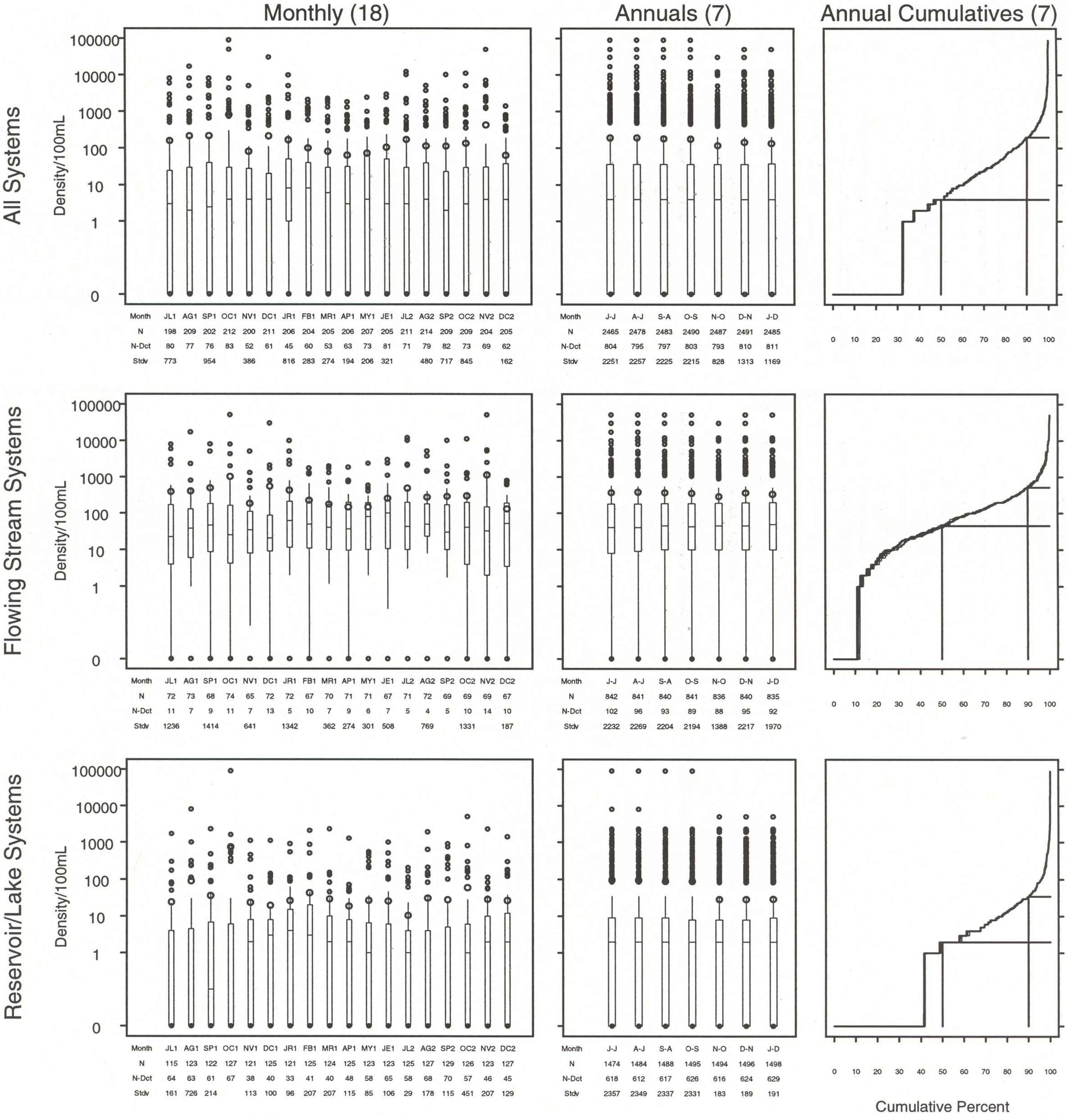
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1419	1416	1410	1404	1402	1407	1405
N-Dct	24	23	23	23	21	21	23
Stdv	17520		12021		17026		22059

Reservoir/Lake Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	192	192	198	196	197	201	197	197	201	199	193	200	188	193	197	199	195	195
N-Dct	39	24	25	24	22	24	21	28	29	28	35	38	30	28	36	33	26	29
Stdv	9020		4688		3871		5696		619	985		5035		14191		18074		2753

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2363	2359	2360	2359	2362	2360	2354
N-Dct	337	328	332	343	352	356	361
Stdv	6996	6648	6571	8380	9303	9380	9354

Exhibit C-12:
Concentration of Fecal Coliform in Plant Influent, ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	198	209	202	212	200	211	206	204	205	206	207	205	211	214	209	209	204	205
N-Dct	80	77	76	83	52	61	45	60	53	63	73	81	71	79	82	73	69	62
Stdv	773		954		386		816	283	274	194	206	321		480	717	845		162

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2465	2478	2483	2490	2487	2491	2485
N-Dct	804	795	797	803	793	810	811
Stdv	2251	2257	2225	2215	828	1313	1169

Flowing Stream Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	72	73	68	74	65	72	72	67	70	71	71	67	71	72	69	69	69	67
N-Dct	11	7	9	11	7	13	5	10	7	9	6	7	5	4	5	10	14	10
Stdv	1236		1414		641		1342		362	274	301	508		769		1331		187

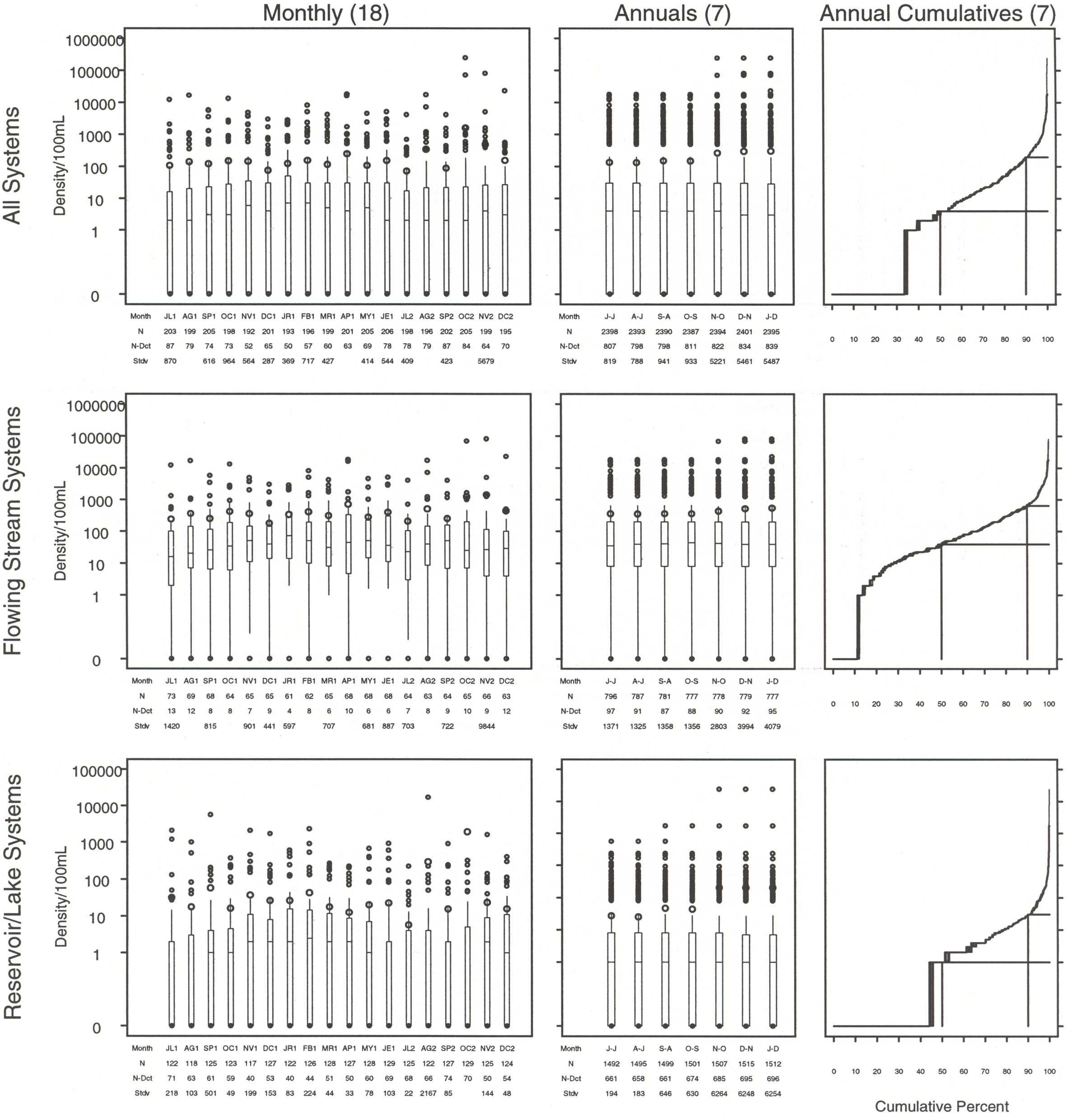
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	842	841	840	841	836	840	835
N-Dct	102	96	93	89	88	95	92
Stdv	2232	2269	2204	2194	1388	2217	1970

Reservoir/Lake Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	115	123	122	127	121	125	121	125	124	125	123	123	125	127	129	126	123	127
N-Dct	64	63	61	67	38	40	33	41	40	48	58	65	58	68	70	57	46	45
Stdv	161	726	214		113	100	96	207	207	115	85	106	29	178	115	451	207	129

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1474	1484	1488	1495	1494	1496	1498
N-Dct	618	612	617	626	616	624	629
Stdv	2357	2349	2337	2331	183	189	191

Exhibit C-13:
Concentration of E. coli in Plant Influent, ICR Plant-Month Data for 18 Months
By Surface Water Category (All, Flowing Streams, Reservoir Lakes)
Monthly & Annual Boxplots, Annual Cumulative Distributions
July 1997 - December 1998



All Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	203	199	205	198	192	201	193	196	199	201	205	206	198	196	202	205	199	195
N-Dct	87	79	74	73	52	65	50	57	60	63	69	78	78	79	87	84	64	70
Stdv	870		616	964	564	287	369	717	427		414	544	409		423		5679	

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	2398	2393	2390	2387	2394	2401	2395
N-Dct	807	798	798	811	822	834	839
Stdv	819	788	941	933	5221	5461	5487

Flowing Stream Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	73	69	68	64	65	65	61	62	65	68	68	68	64	63	64	65	66	63
N-Dct	13	12	8	8	7	9	4	8	6	10	6	6	7	8	9	10	9	12
Stdv	1420		815		901	441	597		707		681	887	703		722		9844	

Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	796	787	781	777	778	779	777
N-Dct	97	91	87	88	90	92	95
Stdv	1371	1325	1358	1356	2803	3994	4079

Reservoir/Lake Systems

Month	JL1	AG1	SP1	OC1	NV1	DC1	JR1	FB1	MR1	AP1	MY1	JE1	JL2	AG2	SP2	OC2	NV2	DC2
N	122	118	125	123	117	127	122	126	128	127	128	129	125	122	127	129	125	124
N-Dct	71	63	61	59	40	53	40	44	51	50	60	69	68	66	74	70	50	54
Stdv	218	103	501	49	199	153	83	224	44	33	78	103	22	2167	85		144	48

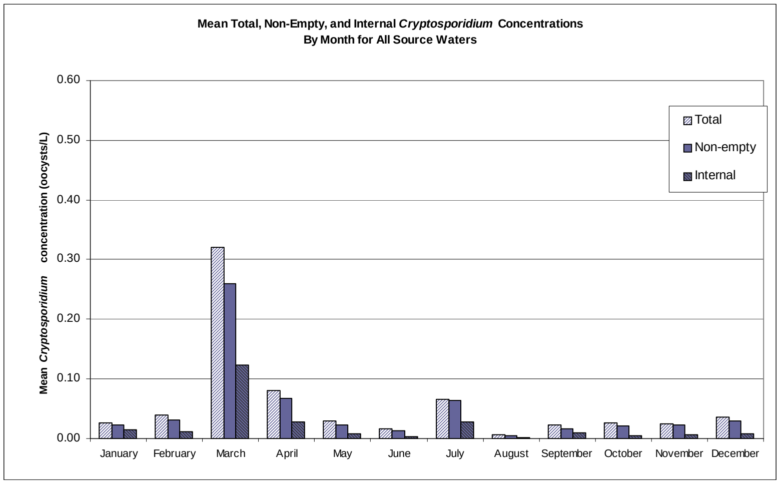
Month	J-J	A-J	S-A	O-S	N-O	D-N	J-D
N	1492	1495	1499	1501	1507	1515	1512
N-Dct	661	658	661	674	685	695	696
Stdv	194	183	646	630	6264	6248	6254

Appendix D. Graphs of Observed Supplemental Survey Data

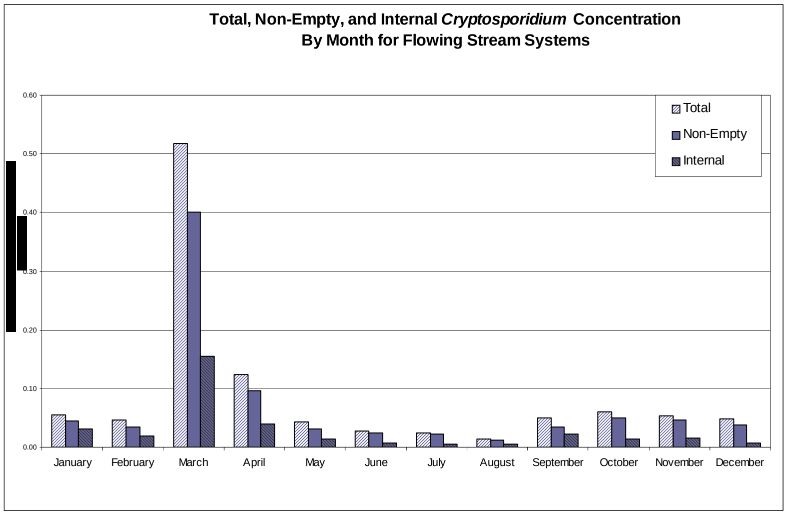
D.1 Exhibit List

Exhibit	**Title**
D.2	Monthly Mean *Cryptosporidium* Concentration by Source Water Type
D.3	Monthly Mean *Giardia* Concentration by Source Water Type
D.4	Monthly Mean Total Coliform Concentration by Source Water Type
D.5	Monthly Mean Fecal Coliform Concentration by Source Water Type
D.6	Monthly Mean *E. coli* Concentration by Source Water Type

Exhibit D.2 ICR Supplemental Surveys Monthly Mean *Cryptosporidium* Concentration by Source Water Type

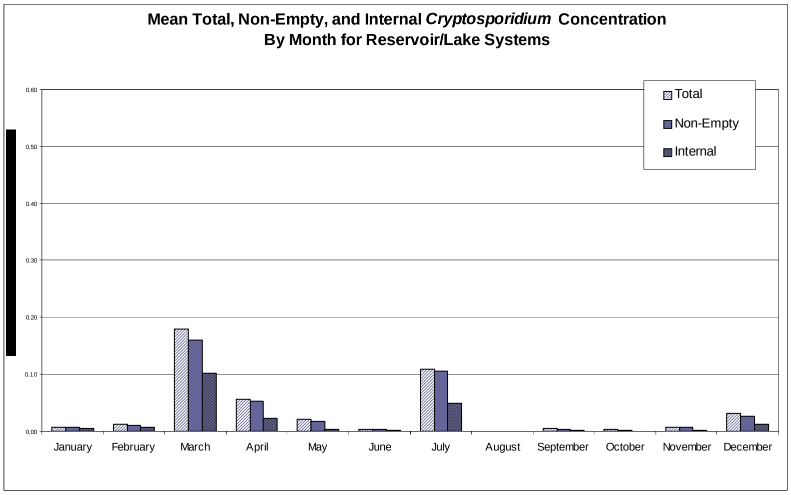


Mean Total, Non-Empty, and Internal *Cryptosporidium* Concentrations
By Month for All Source Waters



Total, Non-Empty, and Internal *Cryptosporidium* Concentration
By Month for Flowing Stream Systems

[redacted]



Mean Total, Non-Empty, and Internal *Cryptosporidium* Concentration
By Month for Reservoir/Lake Systems

[redacted]

Exhibit D.3 ICR Supplemental Surveys Monthly Mean *Giardia* Concentration by Source Water Type

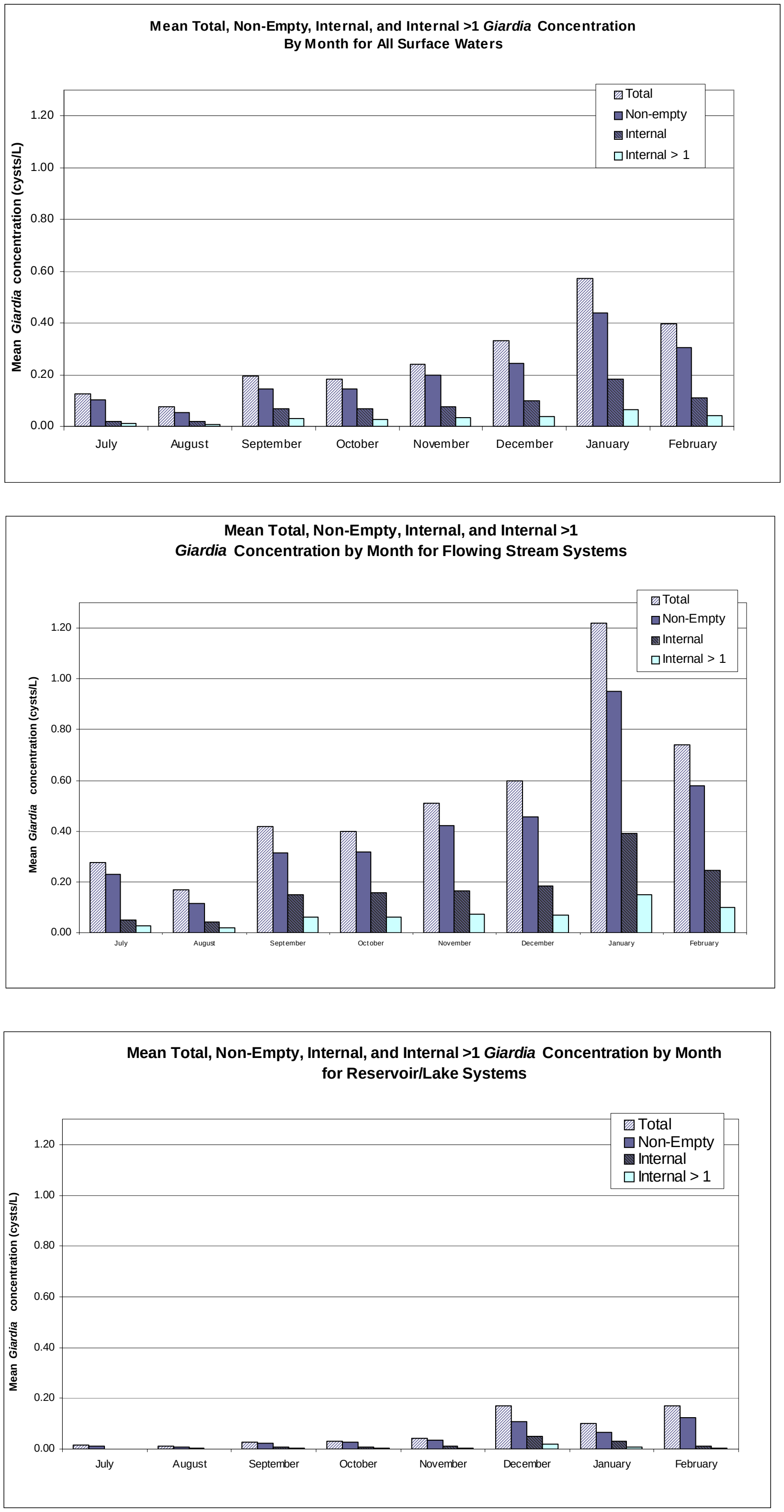
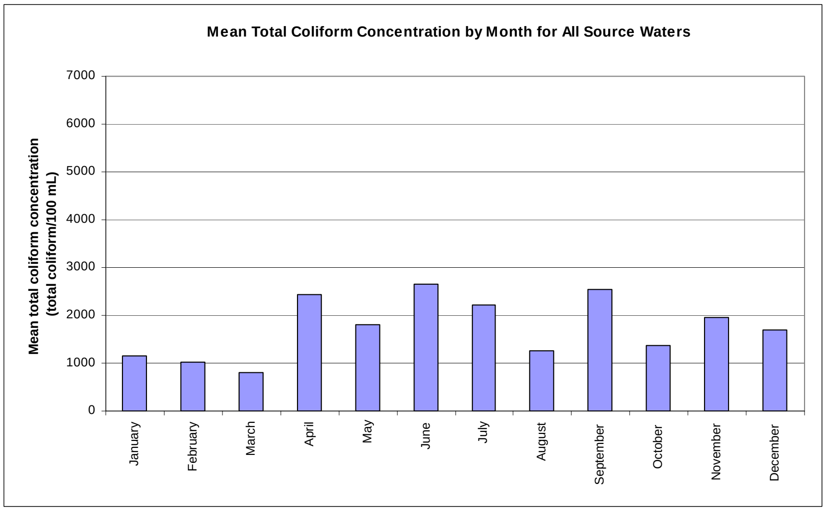
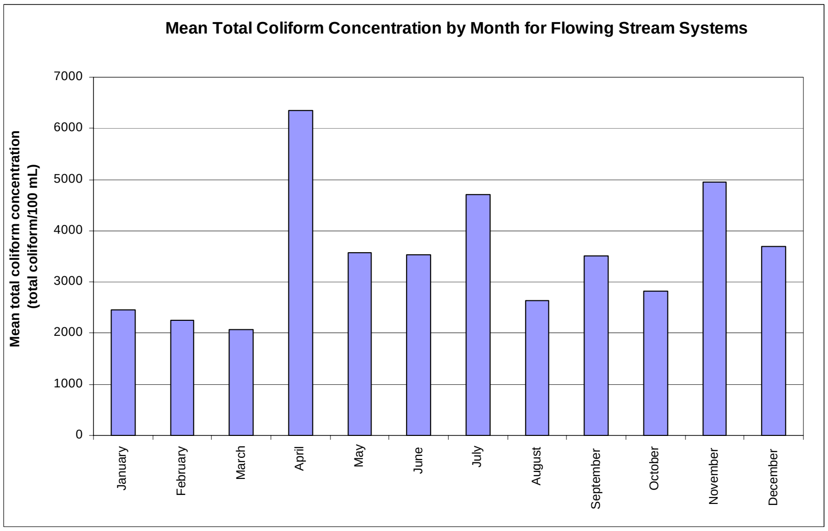


Exhibit D.4 ICR Supplemental Surveys Monthly Mean Total Coliform Concentration by Source Water Type





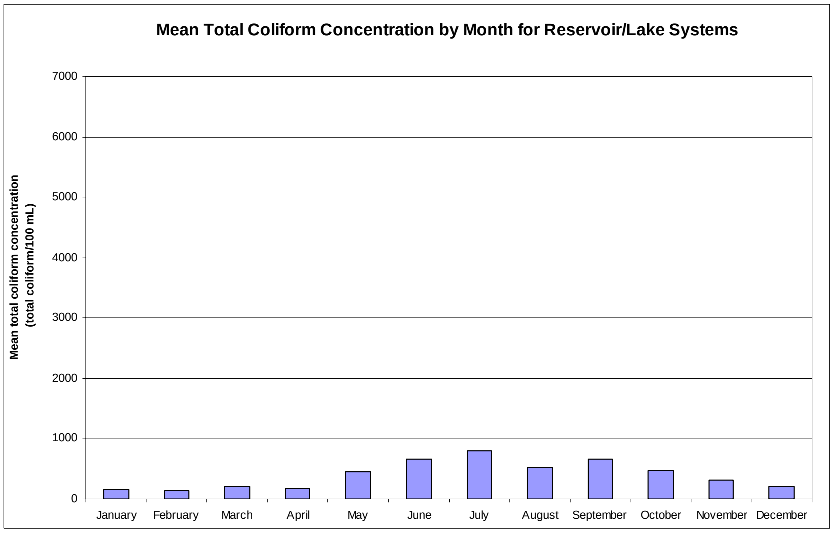
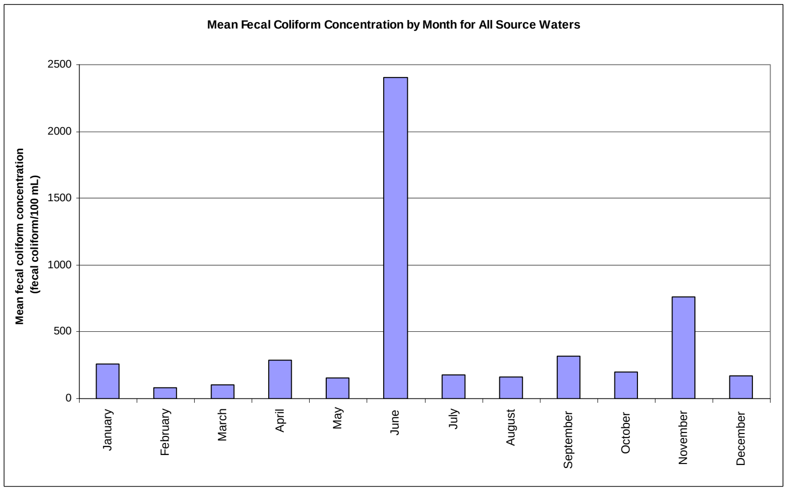
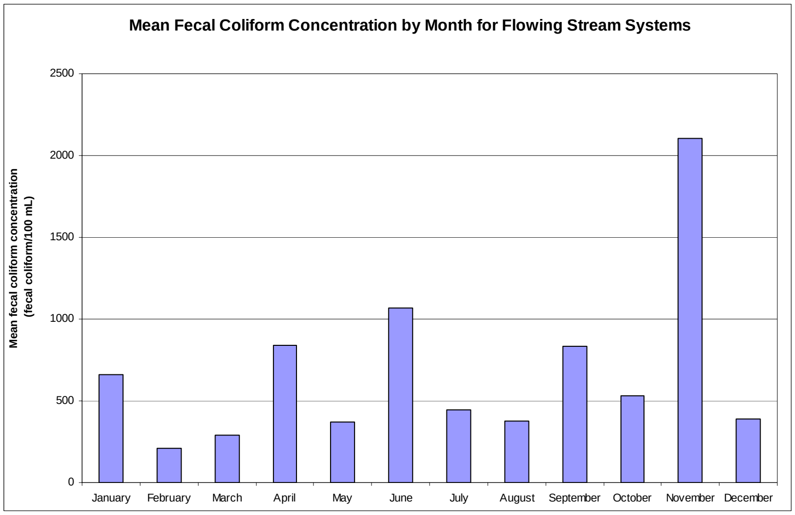


Exhibit D.5 ICR Supplemental Surveys Monthly Mean Fecal Coliform Concentration by Source Water Type





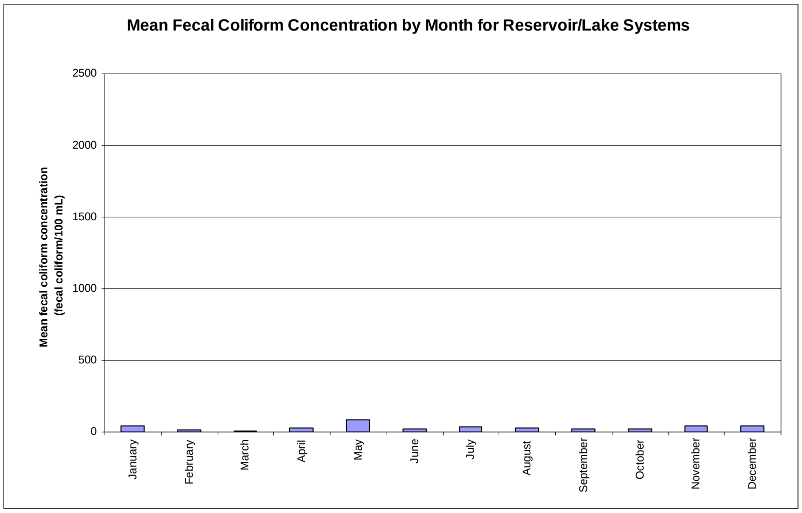
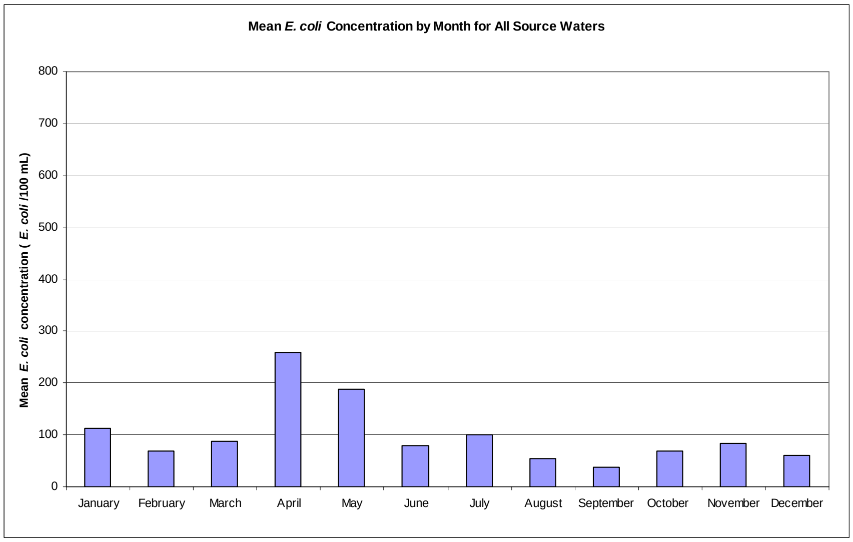
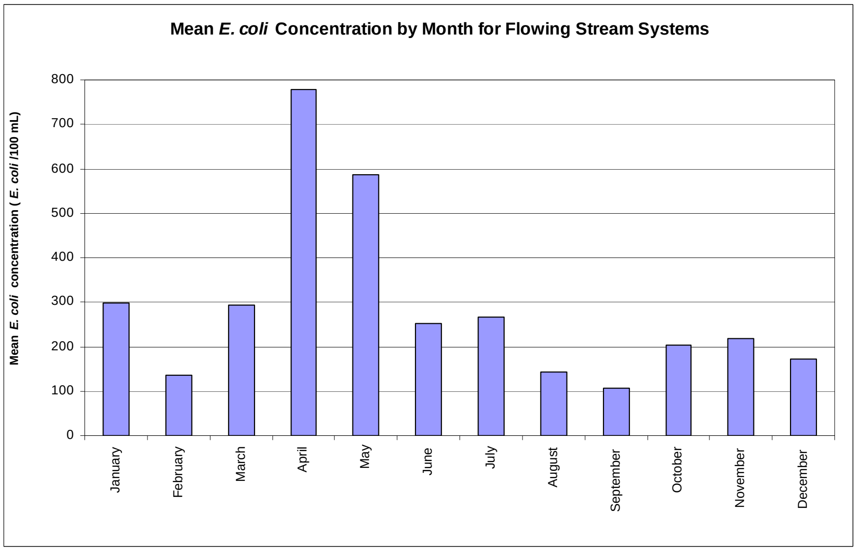
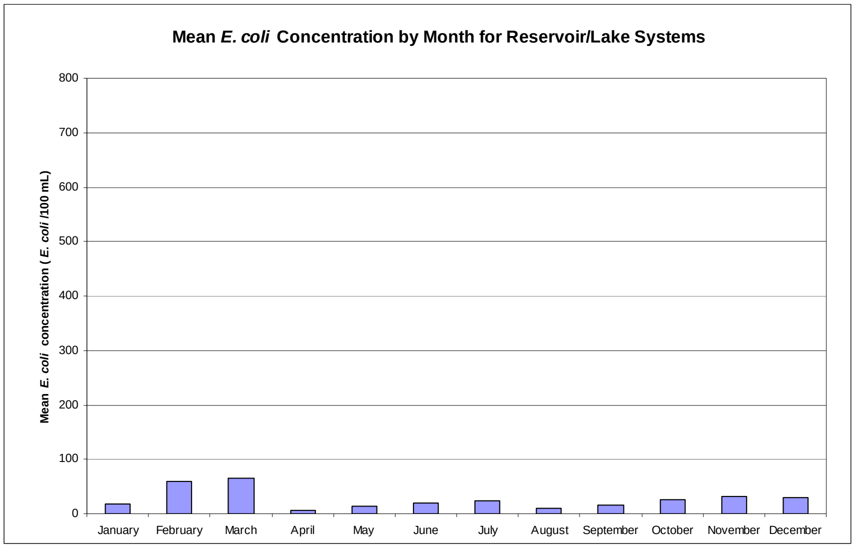


Exhibit D.6 ICR Supplemental Surveys Monthly Mean *E. coli* Concentration by Source Water Type







Appendix E. Bayesian Analysis Cumulative Distribution Functions

Exhibit E.1 Table of Graphs

Exhibit Number	Data Source	Source Water	Pathogen
2	ICR	All	Crypto-Total
3	ICR	All	Crypto-Non-Empty
4	ICR	All	Crypto-Internal Structure
5	ICR	Flowing Stream	Crypto-Total
6	ICR	Flowing Stream	Crypto-Non-Empty
7	ICR	Flowing Stream	Crypto-Internal Structure
8	ICR	Reservoir/Lake	Crypto-Total
9	ICR	RL	Crypto-Non-Empty
10	ICR	RL	Crypto-Internal Structure
11	ICR	All	Giardia-Total
12	ICR	All	Giardia-Non-Empty
13	ICR	All	Giardia-Internal Structure
14	ICR	Flowing Stream	Giardia-Total
15	ICR	Flowing Stream	Giardia-Non-Empty
16	ICR	Flowing Stream	Giardia-Internal Structure
17	ICR	RL	Giardia-Total
18	ICR	RL	Giardia-Non-Empty
19	ICR	RL	Giardia-Internal Structure
20	Supplemental Survey - Large Plants	All	Crypto-Total
21	Supplemental Survey - Medium Plants	All	Crypto-Total
22	Supplemental Survey - Large Plants	All	Crypto-Non-Empty
23	Supplemental Survey - Medium Plants	All	Crypto-Non-Empty
24	Supplemental Survey - Large Plants	All	Crypto-Internal Structure
25	Supplemental Survey - Medium Plants	All	Crypto-Internal Structure

Exhibit Number	**Data Source**	**Source Water**	**Pathogen**
26	Supplemental Survey - Large Plants	Flowing Stream	Crypto-Total
27	Supplemental Survey - Medium Plants	Flowing Stream	Crypto-Total
28	Supplemental Survey - Large Plants	Flowing Stream	Crypto-Non-Empty
29	Supplemental Survey - Medium Plants	Flowing Stream	Crypto-Non-Empty
30	Supplemental Survey - Large Plants	Flowing Stream	Crypto-Internal Structure
31	Supplemental Survey - Medium Plants	Flowing Stream	Crypto-Internal Structure
32	Supplemental Survey - Large Plants	Reservoir/Lake	Crypto-Total
33	Supplemental Survey - Medium Plants	Reservoir/Lake	Crypto-Total
34	Supplemental Survey - Large Plants	Reservoir/Lake	Crypto-Non-Empty
35	Supplemental Survey - Medium Plants	Reservoir/Lake	Crypto-Non-Empty
36	Supplemental Survey - Large Plants	Reservoir/Lake	Crypto-Internal Structure
37	Supplemental Survey - Medium Plants	Reservoir/Lake	Crypto-Internal Structure

Exhibit E-2 Cumulative Distribution of Total Cryptosporidium Oocysts in All Plants

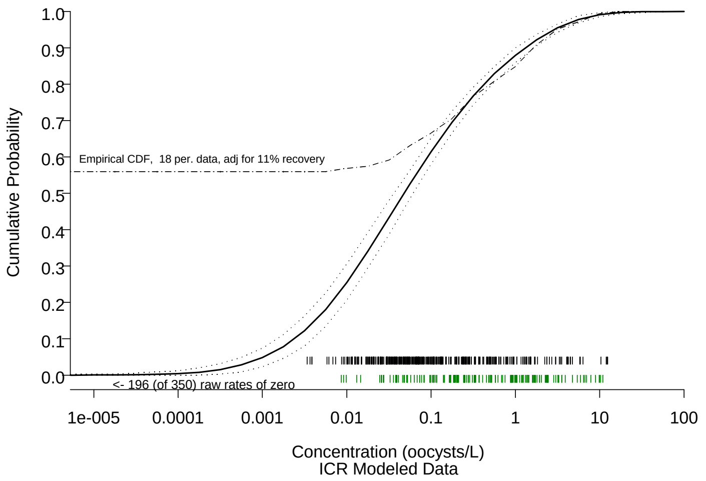


Exhibit E-3 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in All Plants

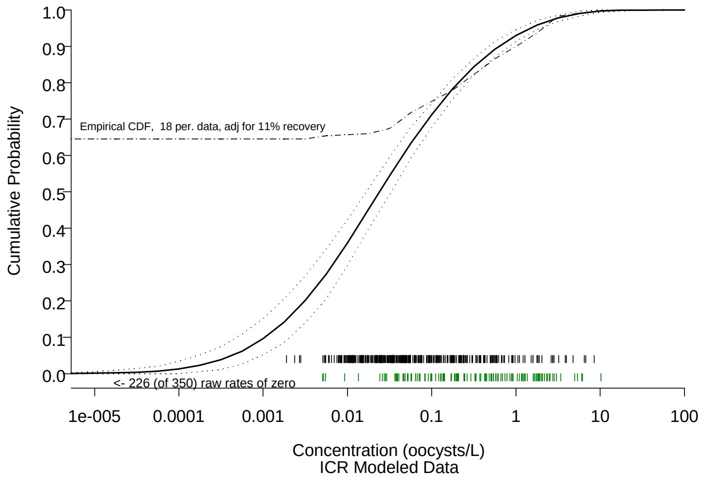


Exhibit E-4 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in All Plants

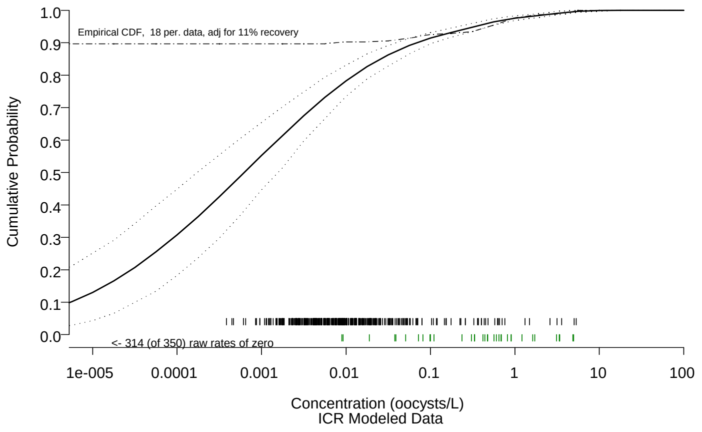


Exhibit E-5 Cumulative Distribution of Total Cryptosporidium Oocysts in Flowing Stream Plants

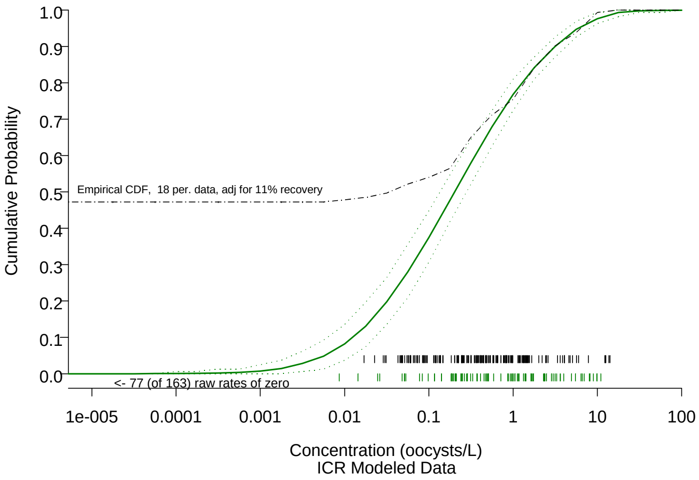


Exhibit E-6 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in Flowing Stream Plants

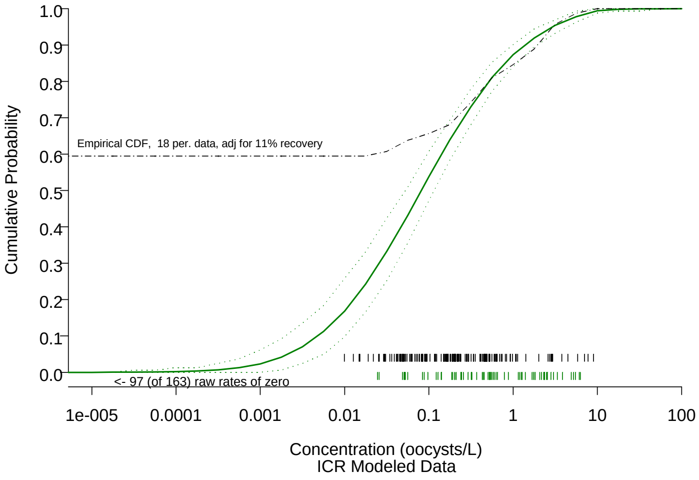


Exhibit E-7 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in Flowing Stream Plants

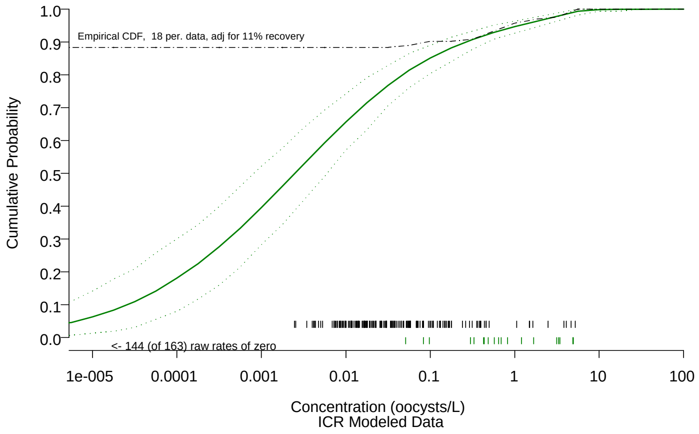


Exhibit E-8 Cumulative Distribution of Total Cryptosporidium Oocysts in Reservoir/Lake Plants

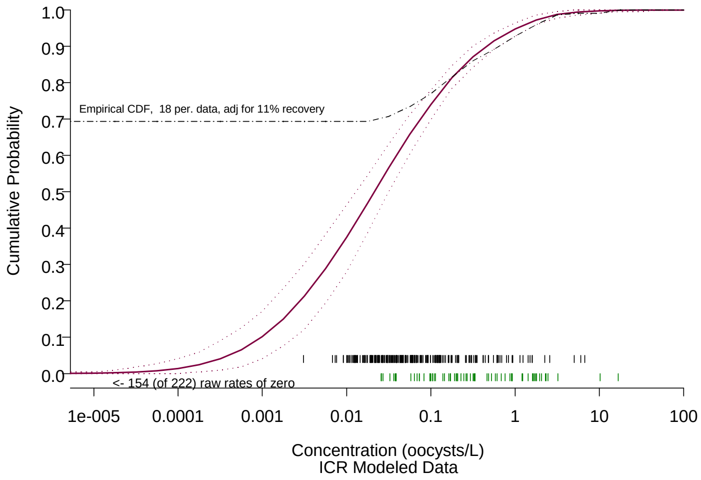


Exhibit E-9 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in RL Plants

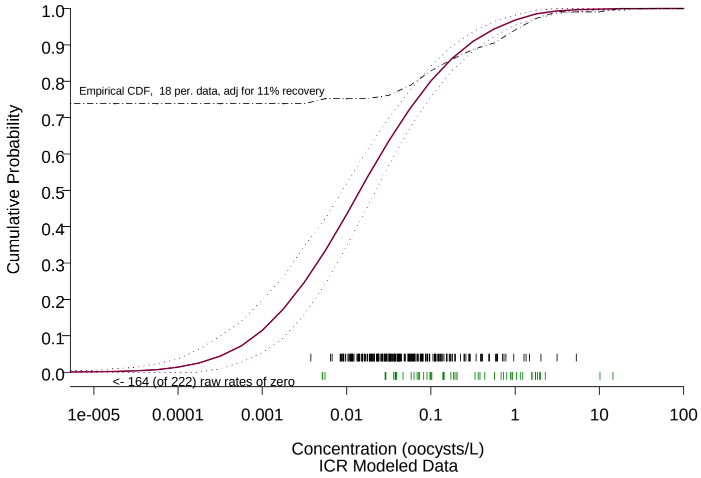


Exhibit E-10 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in RL Plants

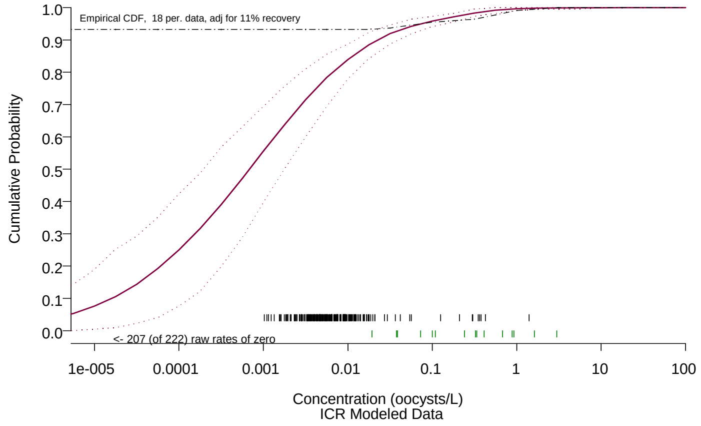


Exhibit E-11 Cumulative Distribution of Total Giardia Cysts in All Plants

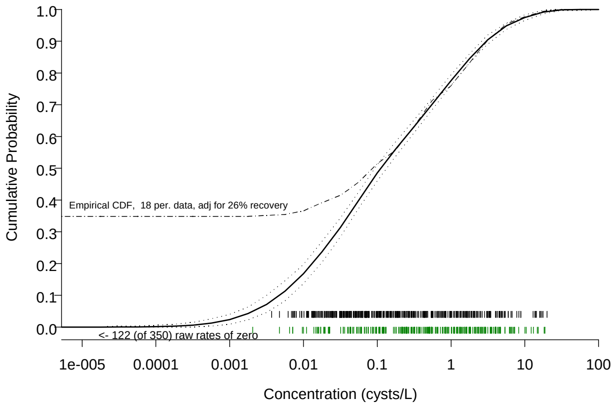


Exhibit E-12 Cumulative Distribution of Non-Empty Giardia Cysts in All Plants

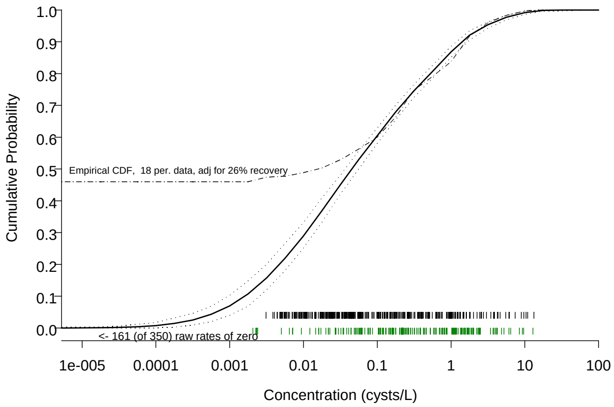


Exhibit E-13 Cumulative Distribution of Giardia Cysts with Internal Structures in All Plants

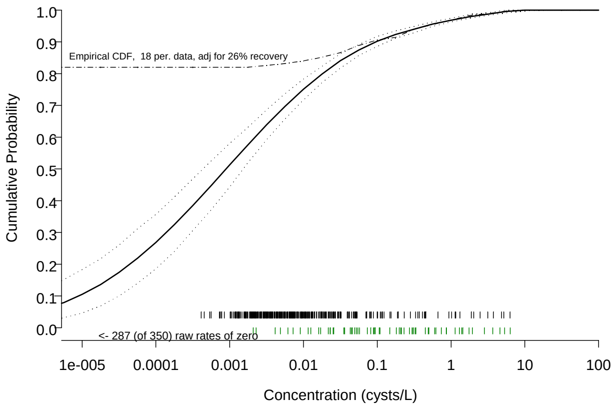


Exhibit E-14 Cumulative Distribution of Total Giardia Cysts in Flowing Stream Plants

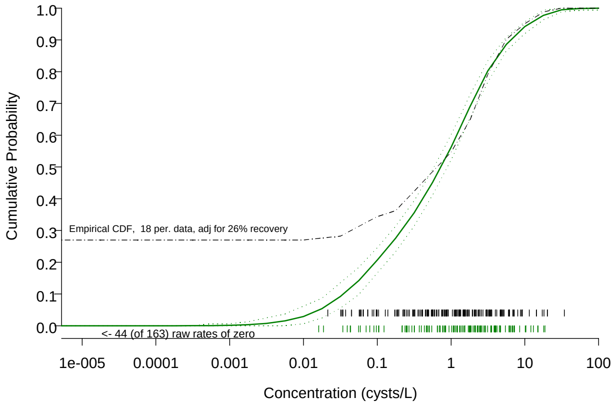


Exhibit E-15 Cumulative Distribution of Non-Empty Giardia Cysts in Flowing Stream Plants

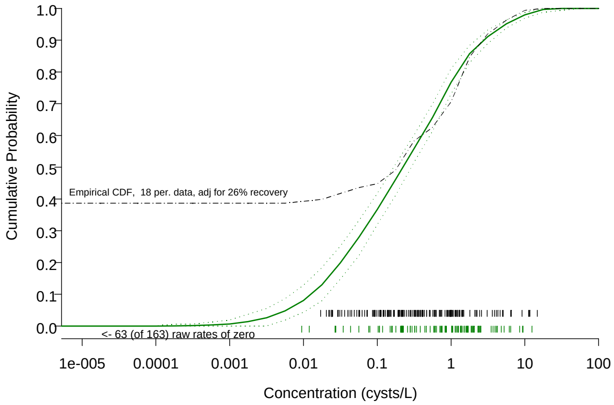


Exhibit E-16 Cumulative Distribution of Giardia Cysts with Internal Structures in Flowing Stream Plants

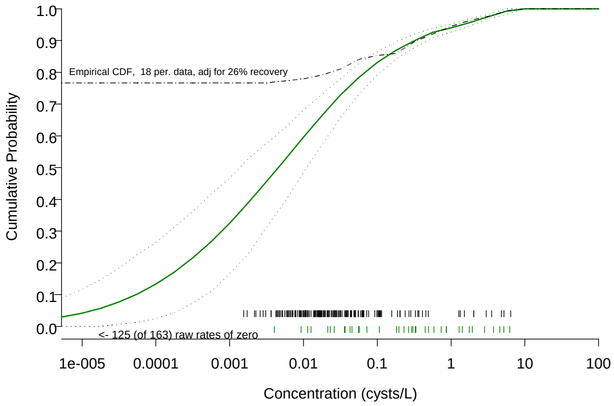


Exhibit E-17 Cumulative Distribution of Total Giardia Cysts in Reservoir/Lake Plants

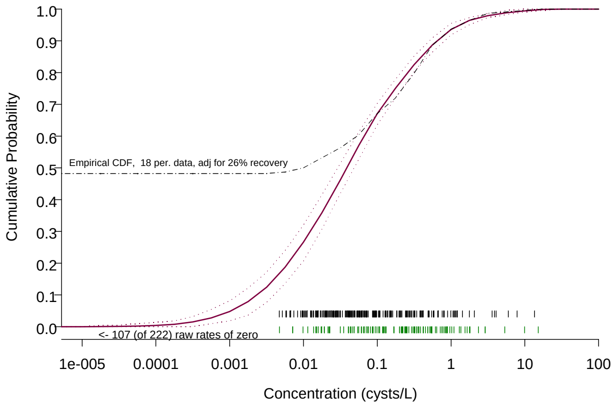


Exhibit E-18 Cumulative Distribution of Non-Empty Giardia Cysts in RL Plants

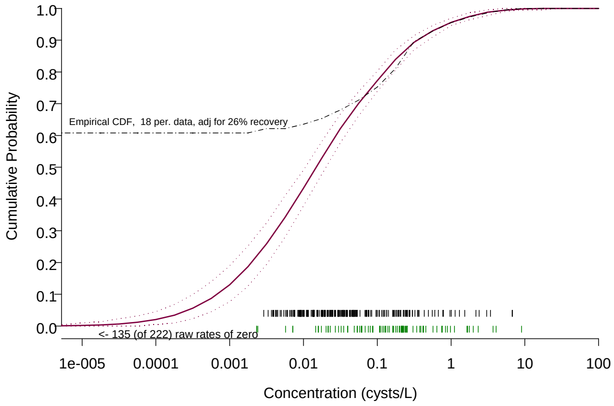


Exhibit E-19 Cumulative Distribution of Giardia Cysts with Internal Structures in RL Plants

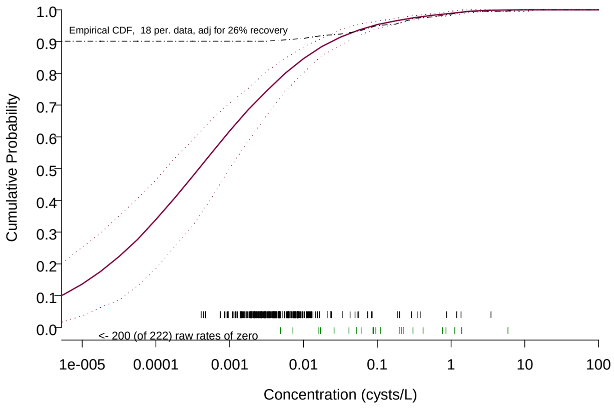


Exhibit E-20 Cumulative Distribution of Total Cryptosporidium Oocysts in All Plants

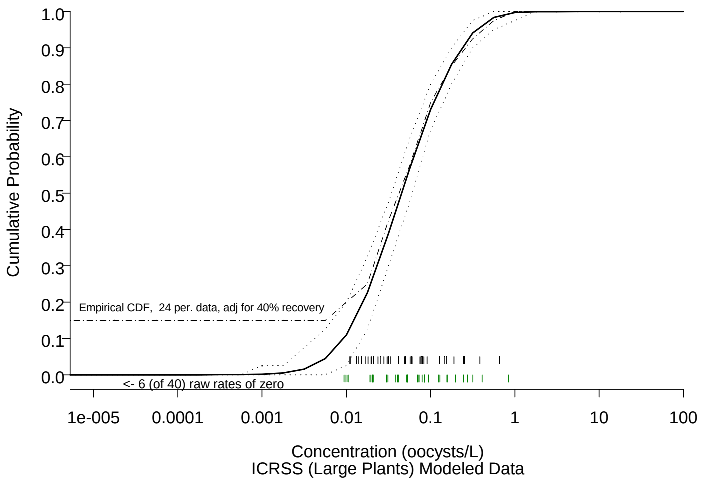


Exhibit E-21 Cumulative Distribution of Total Cryptosporidium Oocysts in All Plants

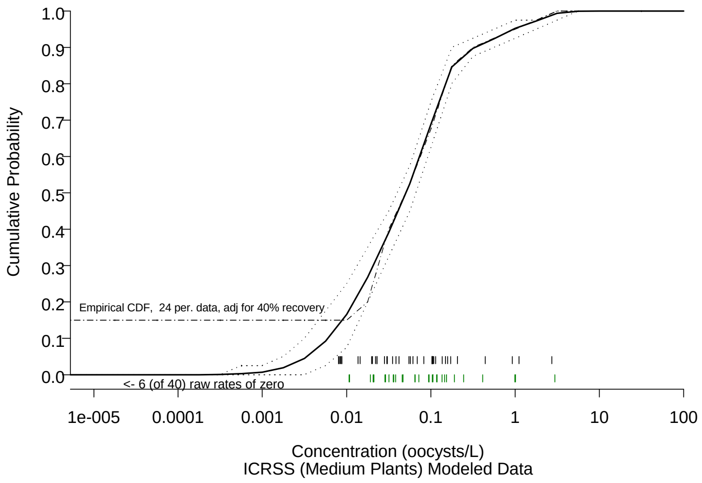


Exhibit E-22 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in All Plants

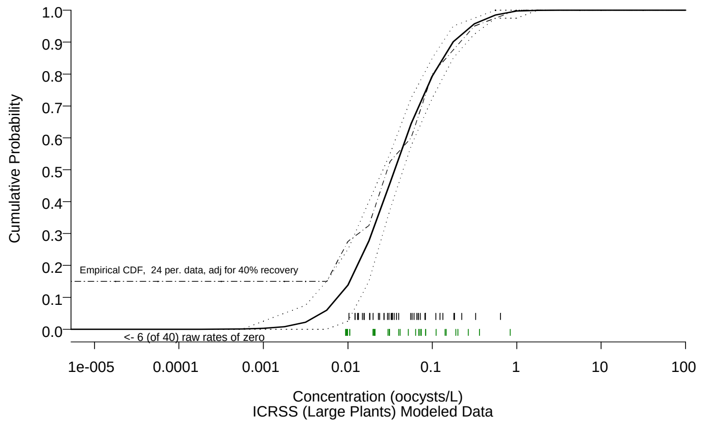


Exhibit E-23 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in All Plants

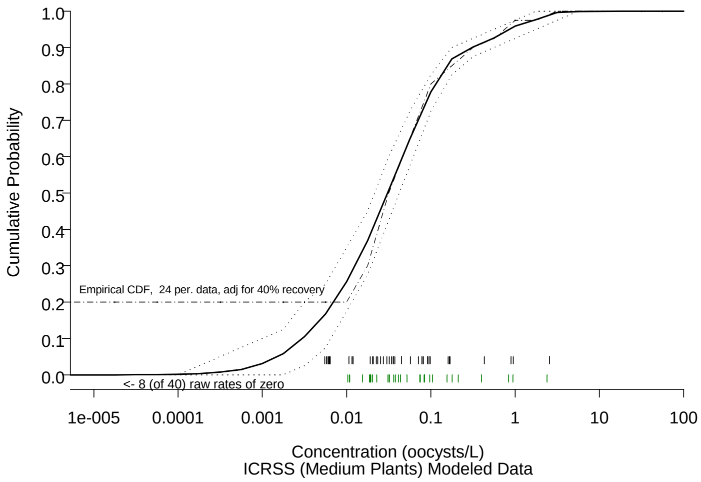


Exhibit E-24 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in All Plants

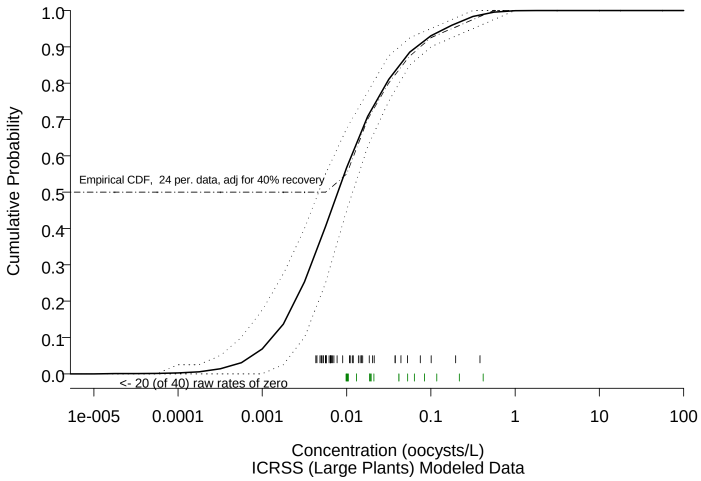


Exhibit E-25 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in All Plants

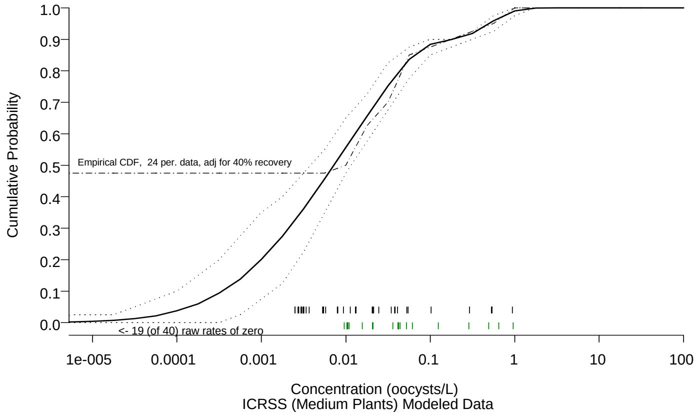


Exhibit E-26 Cumulative Distribution of Total Cryptosporidium Oocysts in Flowing Stream Plants

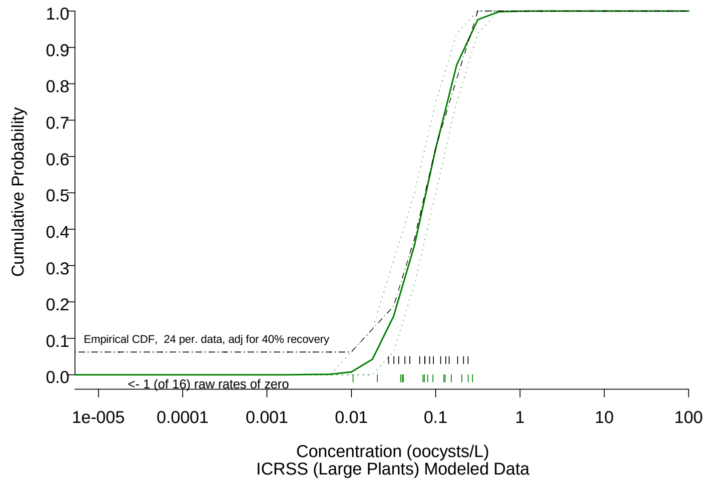


Exhibit E-27 Cumulative Distribution of Total Cryptosporidium Oocysts in Flowing Stream Plants

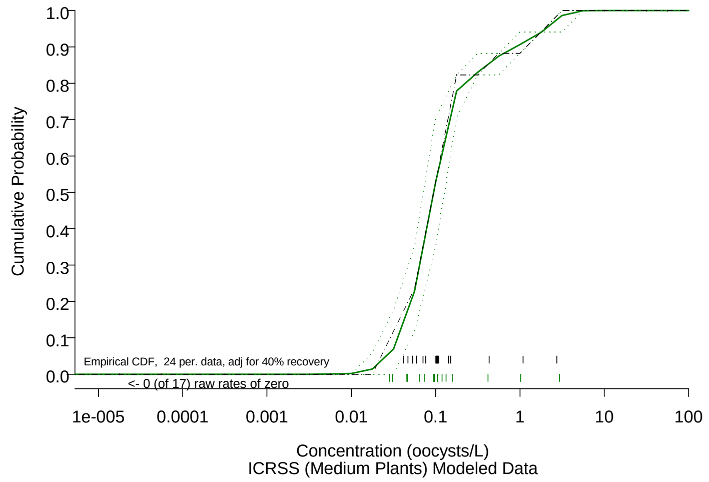


Exhibit E-28 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in Flowing Stream Plants

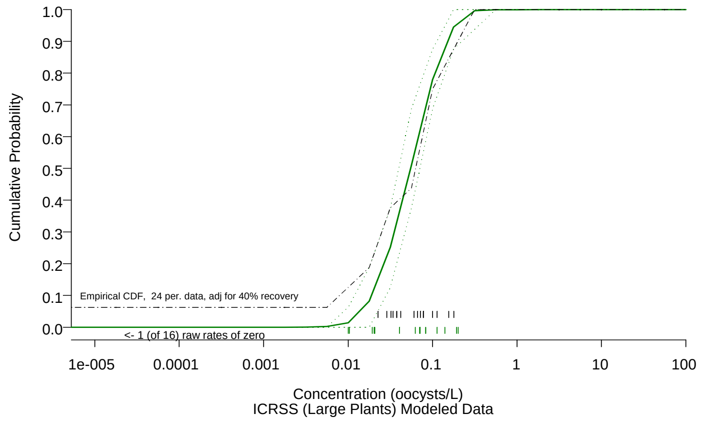


Exhibit E-29 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in Flowing Stream Plants

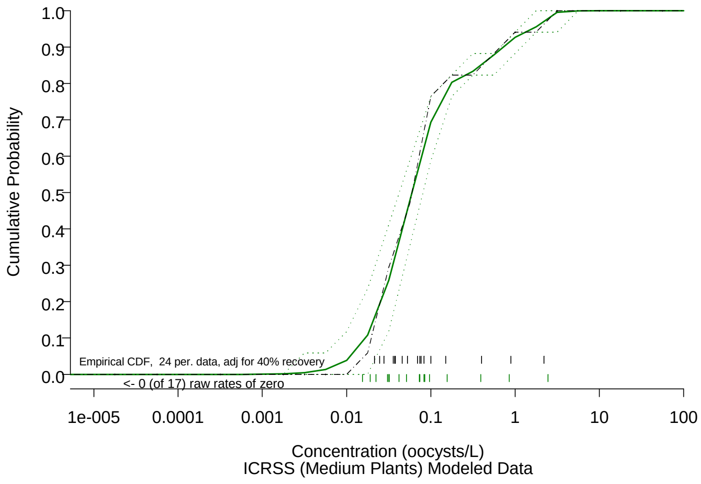


Exhibit E-30 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in Flowing Stream Plants

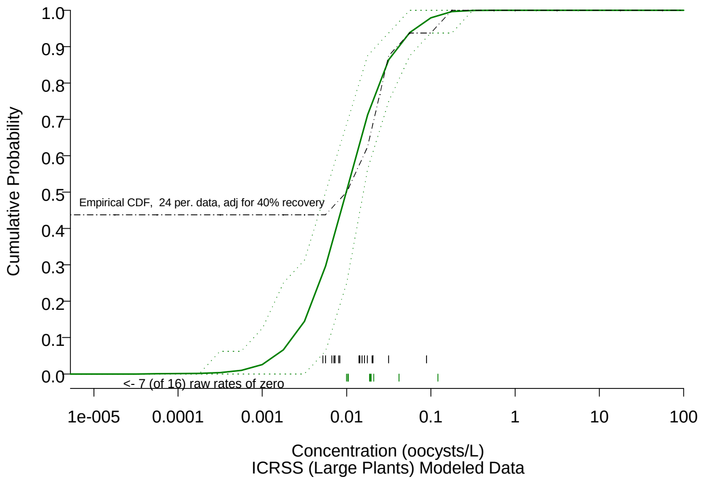


Exhibit E-31 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in Flowing Stream Plants

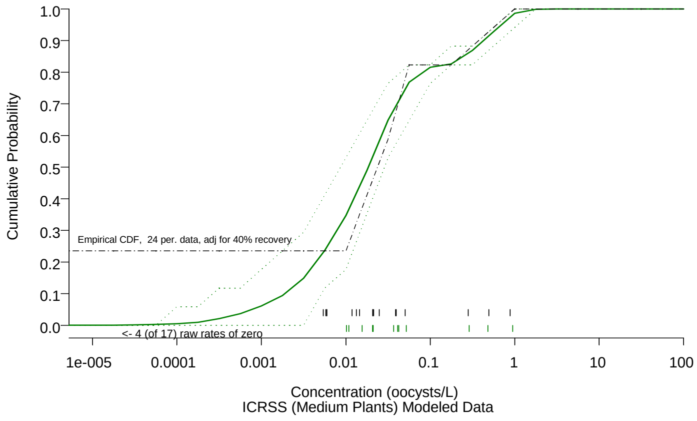


Exhibit E-32 Cumulative Distribution of Total Cryptosporidium Oocysts in Reservoir/Lake Plants

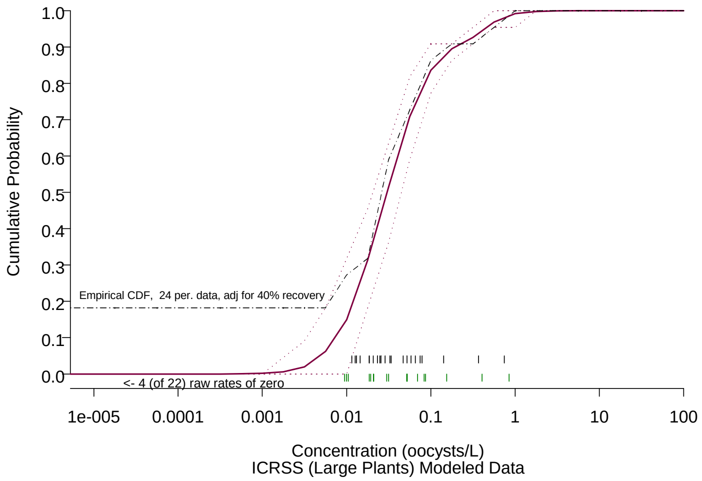


Exhibit E-33 Cumulative Distribution of Total Cryptosporidium Oocysts in Reservoir/Lake Plants

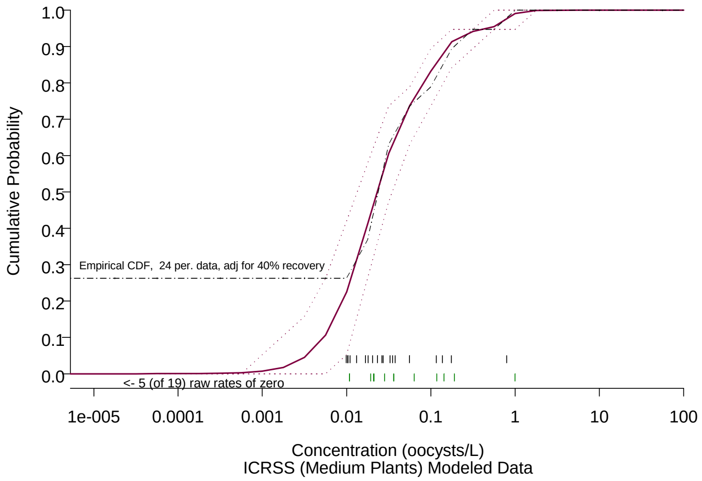


Exhibit E-34 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in Reservoir/Lake Plants

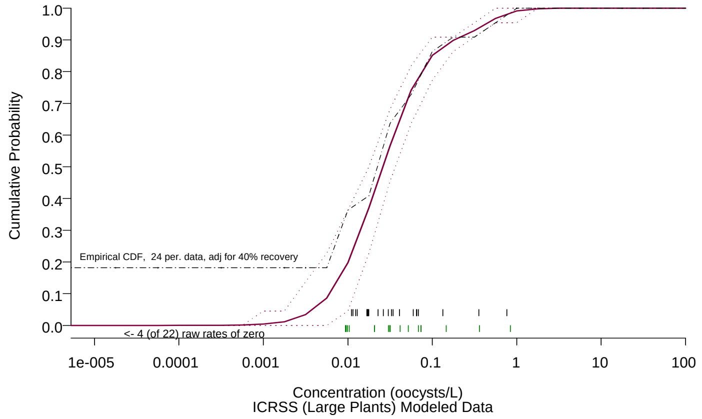


Exhibit E-35 Cumulative Distribution of Non-Empty Cryptosporidium Oocysts in Reservoir/Lake Plants

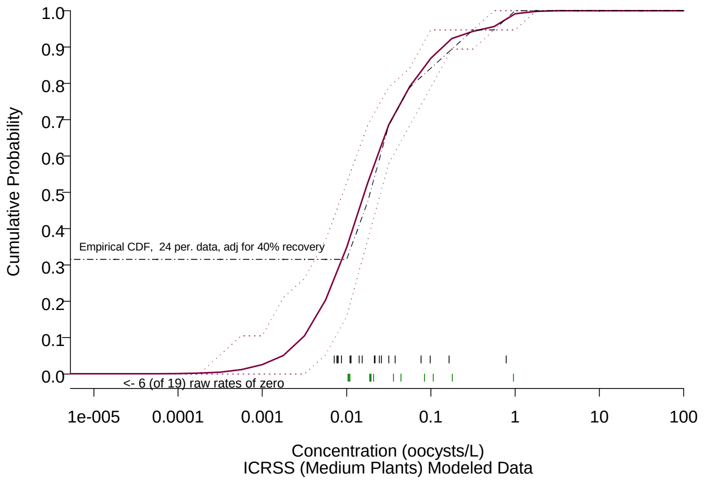


Exhibit E-36 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in Reservoir/Lake Plants

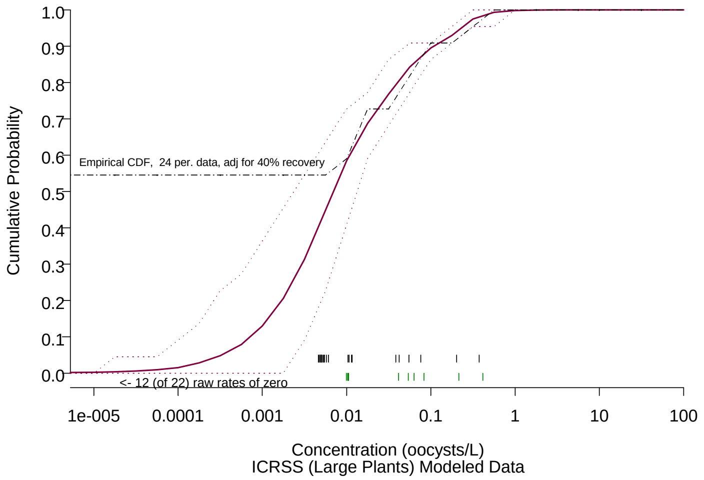
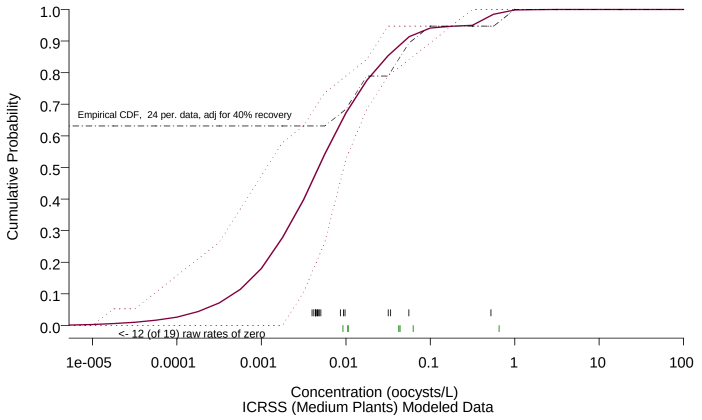


Exhibit E-37 Cumulative Distribution of Cryptosporidium Oocysts with Internal Structures in Reservoir/Lake Plants


Cumulative Probability
Empirical CDF, 24 per. data, adj for 40% recovery
<- 12 (of 19) raw rates of zero
1e-005
Concentration (oocysts/L)
ICRSS (Medium Plants) Modeled Data