Appendices

Page intentionally left blank

Appendix A
Adapted Regression on Order Statistics Methodology

Page intentionally left blank

The method used to fit a censored lognormal distribution and to fill in the censored observations (i.e., values below the instrument reporting limit) was an adaptation and revision of the Regression on Order Statistics (ROS) method as applied in Helsel and Cohn (1988)[20] based on the work in Gilliom and Helsel (1986).[21] The difference between this adapted ROS, and the ROS developed by Helsel and Cohn, is described below. The object is to estimate means and variances for systems where large portions of the data sampled (samples or POE, depending on the data on which the adapted ROS is used) are known to be below some value (e.g. a reporting limit or detection limit). This is carried out by imposing a broad ordering on the data and plotting the complement of an empirical cumulative distribution on log coordinates so that standard regression techniques can be applied to the graph.

The adapted ROS method can be described as follows. Suppose that there are m different reporting limits for the censored values, R_1, R_2, R_3, . . . R_m, arranged in increasing order, and also set R_0 equal to 0 and R_{m+1} equal to ∞. Suppose there are A_0 uncensored values less than R_1, A_1 uncensored values less than R_2 but greater than or equal to R_1, A_2 uncensored values less than R_3 but greater than or equal to R_2, ..., and, in general, A_j uncensored values less than R_{j+1} but greater than or equal to R_j. Also, suppose there are B_j censored or uncensored values below the j^{th} reporting limit, R_j, i.e., either a detected concentration less than R_j or a non-detect with a reporting limit less than or equal to R_j. If the $j+1^{th}$ reporting limit is exceeded, then obviously the j^{th} reporting limit is also exceeded. If the $j+1^{th}$ reporting limit is not exceeded, then an estimate of the probability of exceeding the j^{th} reporting limit is A_j divided by $A_j + B_j$. This estimate is obtained by considering that $A_j + B_j$ values are known to be below the $j+1^{th}$ reporting limit, of which A_j are uncensored values between the two limits and B_j are uncensored or censored values known to be below the lower limit, R_j. (The censored values with reporting limit R_{j+1} cannot be used for this estimate because it is unknown whether those values would have been above or below R_j, had an instrument with this reporting limit been used instead.). This gives the empirical formula for the probability of exceeding the j^{th} reporting limit:

$$p_j = p_{j+1} + [A_j / (A_j + B_j)] (1 - p_{j+1}). \qquad A\text{-}1$$

This equation is solved iteratively, starting with $p_{m+1} = 0$ and letting j = m, m – 1, m – 2, ...

The probability plotting positions (pp) for the A_j uncensored values that are less than R_{j+1} but greater than or equal to R_j are uniformly spread over the probability range $1 - p_j$ to $1 - p_{j+1}$. More precisely, the i^{th} highest of these uncensored values is plotted at probability

$$pp(i) = (1 - p_j) + (p_j - p_{j+1}) \, i / (A_j + 1). \qquad A\text{-}2$$

This is a Weibull based plotting position. Helsel and Cohn (1988) showed that the choice of plotting position did not impact the performance of the estimators examined.

---

[20] Helsel, D. R., and Cohn, T. A. 1988. "Estimation of Descriptive Statistics for Multiply Censored Water Quality Data." *Water Resources Research*, 24, 1997-2004..

[21] Gilliom and Helsel, "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data 1." Estimation Techniques, *Water Resources Research*, 22(2), 135-146, 1986.

To fit the lognormal distribution, a simple linear regression line is fitted to the logarithms of arsenic concentrations (y axis) versus the normal quantiles, defined as G(pp(i)), (x axis), where G is the inverse of the standard normal cumulative distribution function. The intercept and slope are the estimators of the mean, μ, and standard deviation, σ, of the log concentrations.

The ROS method can also be used to substitute values for the censored data. The original ROS method in Helsel and Cohn (1988) chooses plotting points for censored data at reporting limit R_j evenly spread out on the interval from 0 to $1 - p_j$, which is a data-based, non-parametric estimate of the probability of not exceeding the reporting limit. Applying this method to the arsenic data led to some inconsistencies since estimated censored values can exceed the reporting limit. The revised method used for this project avoided this problem by choosing plotting points for censored data evenly spread out on the interval from 0 to the parametrically estimated probability of not exceeding the reporting limit, computed from the fitted lognormal distribution. Thus the probability plotting position for the i^{th} highest of the C_j censored values with reporting limit R_j is

$$pp(i) = \Phi\{(\log R_j - \mu)/\sigma\}\ i / (C_j + 1), \qquad A\text{-}3$$

where Φ is the standard normal cumulative distribution function. The substituted arsenic concentration for that censored value is therefore $\exp\{\mu + \sigma\ G\ [pp(i)]\ \}$, which will always be positive and less than the reporting limit, R_j. After filling the censored observations, the sample mean and standard deviation are calculated using the original uncensored values and the filled-in censored values.

Given a large data set with analytical results that follow a lognormal distribution, the original ROS and the adapted ROS should yield similar results. However, with smaller data sets, the original ROS may yield inconsistent results, in that it predicts that some censored values will exceed the reporting limit. In these cases, the original ROS should slightly overestimate the true distributional parameters. The adaptation of ROS applied in these analyses should correct this bias, and should yield better estimates of the distributional parameters than the original ROS. Thus, the adapted ROS should behave as well as, or better than, the original ROS when applied to the arsenic occurrence data.

An alternative approach that has often been used by researchers to estimate distributions with censored data is the maximum likelihood estimation (MLE) method. This method chooses the fitted distribution to maximizes the likelihood, defined as the product of the fitted probability densities for detected (uncensored) values and the fitted cumulative distribution functions at the detection limit for non-detects (censored values). For lognormal data with a single censoring limit, the MLE method and the ROS method were compared by Kroll and Stedinger (1996).[22] They showed than when the censored data fill-in method was employed with a lognormal MLE, the MLE estimators of the moments and quantiles were more efficient than those of ROS estimators, though for estimators of the mean they were nearly equivalent.

[22] Kroll, C.N. and J.R. Stedinger. 1996. Estimation of moments and quantiles using censored data, *Water Resources Research*, 32(4), 1005-1012.

The crucial issue in the comparison of various estimation methods is the assumed underlying distribution. If the (unknown) true distribution of the water quality data is lognormal, the research by Helsel and Cohn (1988), Kroll and Stedinger (1986), and others shows that the MLE method generally performs best, based on various criteria. If the unknown distribution is quite different from the lognormal distribution, but the lognormal form is fitted, then MLE generally performs worse than ROS, as shown by Helsel and Cohn (1988). This robustness property of ROS stems from the fact that the moments and quantiles are computed using the original uncensored data combined with estimated values for the censored data. Thus the fitting method is only applied to the censored data.

In summary, the adapted ROS method was used to estimate system means if there were at least five detected values (not all equal) and some non-detects. The concentrations for the non-detects were estimated using the adapted ROS rather than the original ROS, which generally leads to lower estimated concentrations that are always below the reporting limit. The system means are then computed by averaging the original detected values and the filled-in non-detect concentrations. At the state level, system means for completely censored systems (with no detected values) were also estimated using the adapted ROS method for use in some of the statistical analyses. However, the state level distributions used for the Regional and national arsenic occurrence analyses in chapter 6 were based on the parametric right-tailed ROS method, which does not use the adapted ROS estimates for such completely censored systems.

This page intentionally left blank

Appendix B
Analysis Results

This page intentionally left blank

Appendix B-1
State Exceedance Probability Distributions

Page intentionally left blank

Table B-1a
Right-tailed ROS State Distributions for Ground Water CWS Systems
Use ROS? = "No" in cases where the substitution method is used instead of ROS due to limited data.

Source Type	State	Fraction of Systems Exceeding Arsenic Concentrations (mg/L) of:										Use ROS?
		2	3	5	10	15	20	25	30	40	50	
GW	AK	0.411968	0.315991	0.211267	0.107406	0.06721	0.04659	0.03438	0.026476	0.017119	0.01196	Yes
GW	AL	0.015111	0.008727	0.004129	0.00135	0.000664	0.000391	0.000256	0.000179	0.000101	6.33E-05	Yes
GW	AR	NA	NA	0.005391	0	0	0	0	0	0	0	No
GW	AZ	NA	NA	0.495376	0.273518	0.171642	0.116469	0.083338	0.061986	0.037257	0.024215	Yes
GW	CA	0.437587	0.325363	0.204616	0.091646	0.051938	0.033169	0.022813	0.016512	0.009601	0.006135	Yes
GW	IL	0.215996	0.156069	0.097806	0.046603	0.028476	0.019543	0.014368	0.011062	0.007186	0.005061	Yes
GW	IN	0.037836	0.017059	0.005387	0.000859	0.000253	9.95E-05	4.64E-05	2.43E-05	8.33E-06	3.5E-06	Yes
GW	KS	0.42795	0.270656	0.124743	0.02969	0.010317	0.004407	0.002149	0.00115	0.0004	0.000166	Yes
GW	KY	NA	NA	0.003926	2.56E-05	5.73E-07	2.6E-08	2E-09	0	0	0	Yes
GW	ME	0.288764	0.216684	0.142597	0.072719	0.046258	0.032669	0.024559	0.019255	0.012872	0.00927	Yes
GW	MI	0.521994	0.41017	0.280041	0.14338	0.088906	0.060836	0.044272	0.033621	0.021162	0.01442	Yes
GW	MN	0.281339	0.203021	0.125408	0.057147	0.033536	0.022216	0.01583	0.011848	0.007327	0.004947	Yes
GW	MO	0.051989	0.034353	0.019445	0.008247	0.004767	0.003164	0.002275	0.001724	0.001097	0.000763	Yes
GW	MT	0.228815	0.150985	0.081222	0.029263	0.014584	0.008503	0.00545	0.003725	0.001979	0.00118	Yes
GW	NC	NA	NA	NA	0.007937	0.001656	0.000467	0.00016	6.27E-05	1.29E-05	3.44E-06	Yes
GW	ND	0.440649	0.344492	0.236891	0.126049	0.081331	0.057721	0.043422	0.033992	0.022582	0.016128	Yes
GW	NH	NA	NA	0.241318	0.122202	0.075744	0.051971	0.037969	0.028962	0.018398	0.012652	Yes
GW	NJ	NA	NA	NA	0.00474	0.00164	0.000722	0.000368	0.000207	7.94E-05	3.64E-05	Yes
GW	NM	0.48692	0.356911	0.215513	0.087183	0.0453	0.026855	0.017303	0.011815	0.006209	0.003639	Yes
GW	NV	NA	NA	0.530386	0.323111	0.220021	0.160044	0.121688	0.095542	0.063066	0.044392	Yes
GW	OH	NA	NA	NA	0.042369	0.016999	0.008169	0.004405	0.002575	0.00104	0.000489	Yes
GW	OK	0.335218	0.234007	0.134824	0.052938	0.027608	0.01658	0.010857	0.007543	0.004105	0.002489	Yes
GW	OR	NA	NA	0.14369	0.056434	0.029344	0.017559	0.011456	0.007931	0.004288	0.002583	Yes
GW	TX	0.272472	0.185434	0.103912	0.039675	0.020479	0.012245	0.008005	0.005561	0.003032	0.001845	Yes
GW	UT	0.327634	0.234033	0.140663	0.060021	0.033344	0.021085	0.014427	0.010416	0.006052	0.003875	Yes

Table B-1b
Right-tailed ROS State Distributions for Surface Water CWS Systems
Use ROS? = "No" in cases where the substitution method is used instead of ROS due to limited data.

Source Type	State	Fraction of Systems Exceeding Arsenic Concentrations (mg/L) of:										Use ROS?
		2	3	5	10	15	20	25	30	40	50	
SW	AK	0.117603	0.079064	0.045129	0.018875	0.01067	0.006926	0.004875	0.003622	0.002223	0.001498	Yes
SW	AL	0.005162	0.002125	0.00062	9.48E-05	2.83E-05	1.14E-05	5.48E-06	2.96E-06	1.08E-06	4.79E-07	Yes
SW	AR	NA	NA	0	0	0	0	0	0	0	0	Yes
SW	AZ	NA	NA	0.307951	0.071823	0.021469	0.007696	0.003145	0.001418	0.000358	0.000111	Yes
SW	CA	0.180545	0.130138	0.081751	0.039496	0.024476	0.017014	0.012654	0.009845	0.006515	0.004663	Yes
SW	IL	0.016466	0.009268	0.004217	0.001292	0.000608	0.000346	0.00022	0.000151	8.1E-05	4.93E-05	Yes
SW	IN	0.058824	0.039216	0	0	0	0	0	0	0	0	Yes
SW	KS	0.097497	0.01924	0.001165	6.29E-06	1.37E-07	6E-09	0	0	0	0	Yes
SW	KY	NA	NA	0.001088	6.4E-07	2E-09	0	0	0	0	0	Yes
SW	ME	0.084409	0.062308	0.041207	0.022225	0.015023	0.011222	0.008878	0.007294	0.005298	0.004101	Yes
SW	MI	0.095131	0.071168	0.047919	0.026535	0.018233	0.013786	0.011013	0.009121	0.006712	0.00525	Yes
SW	MN	2.64E-05	1E-09	0	0	0	0	0	0	0	0	Yes
SW	MO	0.031242	0.014966	0.005245	0.001014	0.000344	0.000151	7.77E-05	4.41E-05	1.74E-05	8.19E-06	Yes
SW	MT	0.264849	0.145369	0.055271	0.00997	0.00293	0.00111	0.000492	0.000244	7.5E-05	2.83E-05	Yes
SW	NC	NA	NA	NA	0.00026	2.61E-06	5E-08	2E-09	0	0	0	Yes
SW	ND	0.087907	0.015543	0.000771	2.84E-06	4.6E-08	2E-09	0	0	0	0	Yes
SW	NH	NA	NA	0.249156	0.005336	0.000131	4.73E-06	2.41E-07	1.6E-08	0	0	Yes
SW	NJ	NA	NA	NA	0.000107	6.65E-06	7.23E-07	1.12E-07	2.2E-08	1E-09	0	Yes
SW	NM	0.167619	0.081423	0.026205	0.003701	0.000935	0.000317	0.000129	5.94E-05	1.63E-05	5.6E-06	Yes
SW	NV	NA	NA	0.122028	0.074862	0.054608	0.043071	0.035551	0.030237	0.023202	0.018743	Yes
SW	OH	NA	NA	NA	0.000652	1.14E-05	3.55E-07	1.7E-08	1E-09	0	0	Yes
SW	OK	0.044236	0.028273	0.015274	0.006029	0.003327	0.002132	0.00149	0.001102	0.000674	0.000454	Yes
SW	OR	NA	NA	0.031042	0.002679	0.000446	0.000106	3.17E-05	1.11E-05	1.89E-06	4.36E-07	Yes
SW	TX	0.162879	0.055778	0.009189	0.000339	3.08E-05	4.53E-06	9.05E-07	2.24E-07	2.1E-08	3E-09	Yes
SW	UT	0.225641	0.156929	0.092252	0.039117	0.021957	0.01408	0.009778	0.007164	0.004281	0.002813	Yes

Table B-1c.
Right-tailed ROS State Distributions for Ground Water NTNCWS Systems
Use ROS? = "No" in cases where the substitution method is used instead of ROS due to limited data.

Source Type	State	Fraction of Systems Exceeding Arsenic Concentrations (mg/L) of:										Use ROS?
		2	3	5	10	15	20	25	30	40	50	
GW	AK	0.545093	0.434398	0.302924	0.160588	0.101952	0.071042	0.052474	0.040359	0.025941	0.017981	Yes
GW	AL[1]	0.015111	0.008727	0.004129	0.00135	0.000664	0.000391	0.000256	0.000179	0.000101	6.33E-05	Yes
GW	AZ	NA	NA	0.348575	0.186261	0.11786	0.081606	0.059837	0.045673	0.028918	0.01976	Yes
GW	CA	0.436018	0.329302	0.213142	0.101059	0.059817	0.039586	0.028064	0.020861	0.012693	0.008424	Yes
GW	IN	0.032322	0.020645	0.011196	0.004482	0.002506	0.001624	0.001147	0.000856	0.000532	0.000364	Yes
GW	KS	0.288119	0.181684	0.088506	0.025679	0.010765	0.005434	0.003076	0.001884	0.000829	0.000422	Yes
GW	MI	0.486995	0.375774	0.250032	0.123049	0.074387	0.049947	0.035799	0.026843	0.01655	0.011093	Yes
GW	MN	0.294147	0.209667	0.126511	0.054996	0.031126	0.020016	0.013905	0.010179	0.00606	0.003964	Yes
GW	NC	NA	NA	NA	0.014378	0.007735	0.004834	0.003299	0.002388	0.001404	0.000913	Yes
GW	ND	0.358634	0.29753	0.228153	0.150441	0.114305	0.092728	0.078187	0.067647	0.053292	0.043914	Yes
GW	NJ	NA	NA	NA	0.021465	0.010546	0.006099	0.003889	0.002649	0.001402	0.000835	Yes
GW	NM	0.706049	0.564137	0.375319	0.166493	0.088744	0.052784	0.03379	0.022812	0.011641	0.006604	Yes
GW	OR	NA	NA	0.113622	0.038317	0.017867	0.009808	0.005953	0.003872	0.001886	0.001042	Yes
GW	TX	0.258485	0.177249	0.100981	0.039972	0.021235	0.013006	0.00868	0.006141	0.003456	0.002159	Yes

This page intentionally left blank

Appendix B-2
Box Plots

This page intentionally left blank

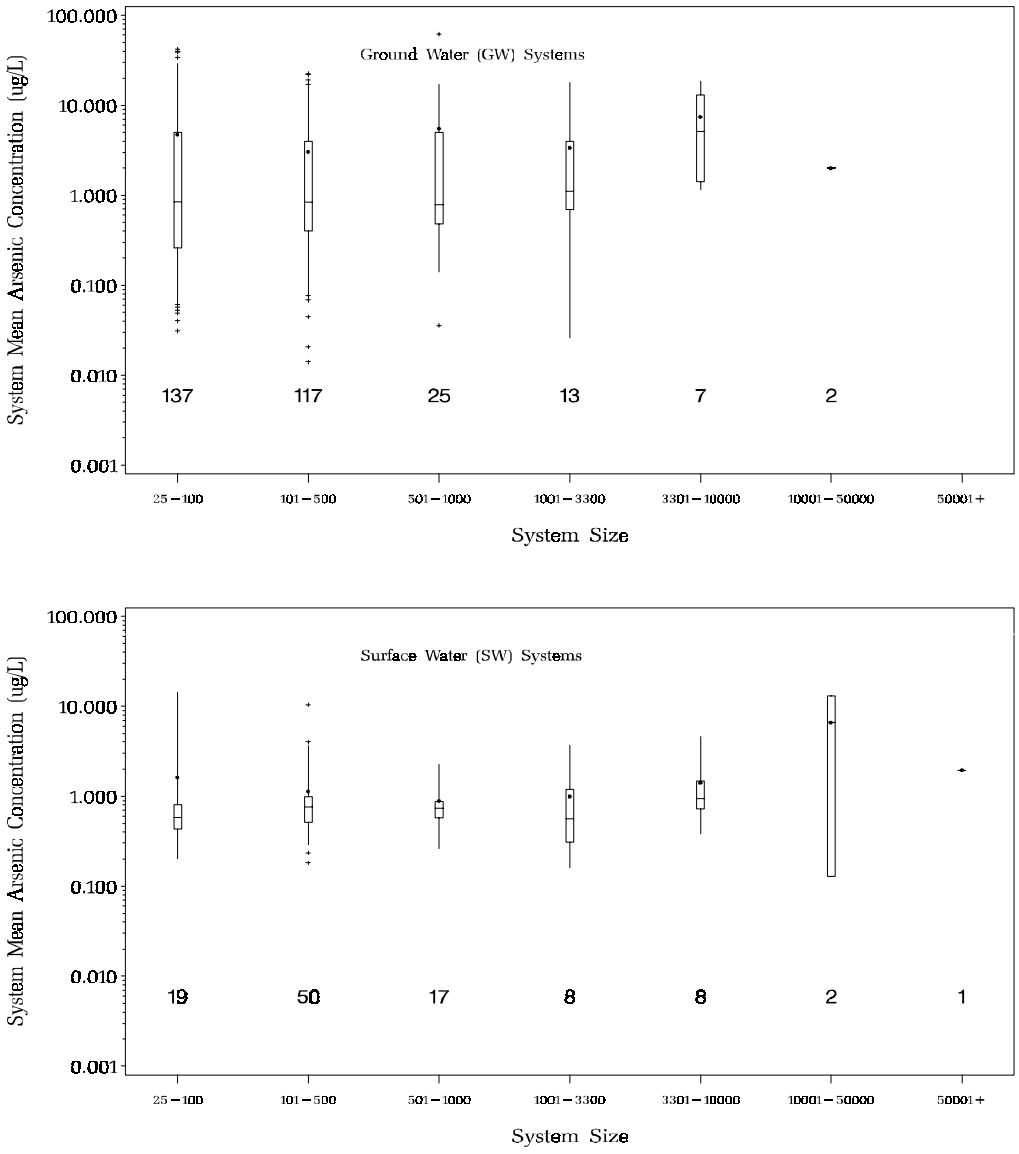
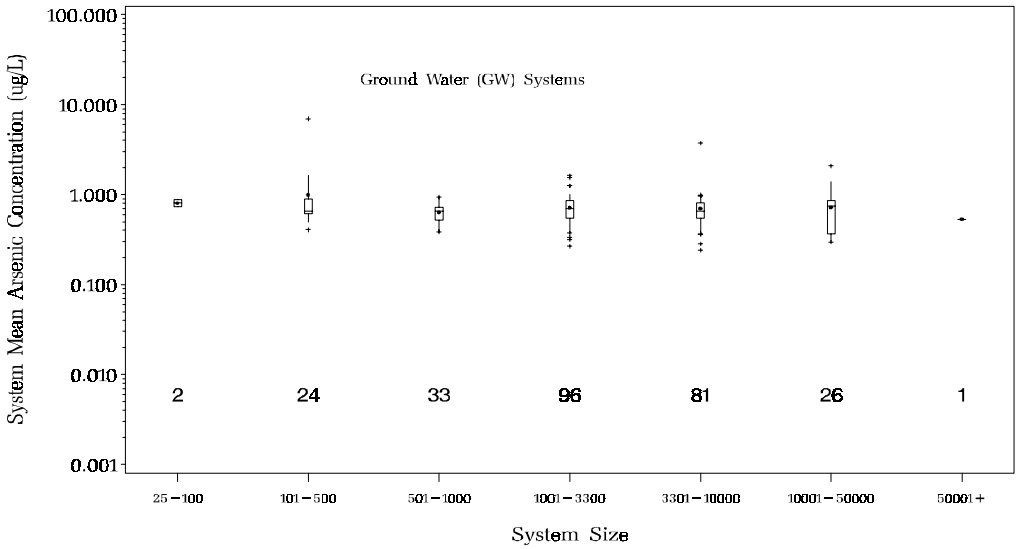


FIGURE 1A. Boxplots of System Means by System Size for
Community Water Systems in State AK
Number of Systems Indicated Below Boxplot



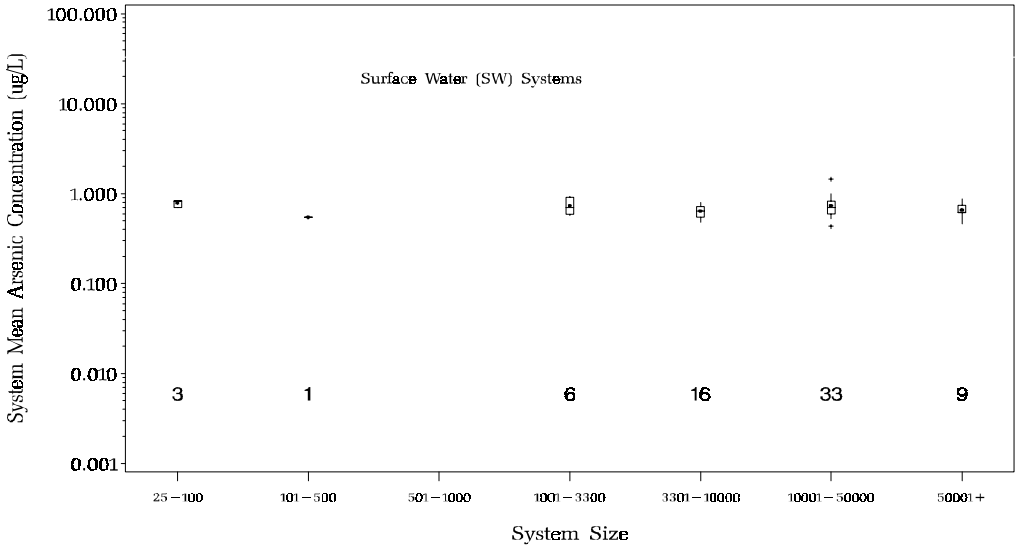
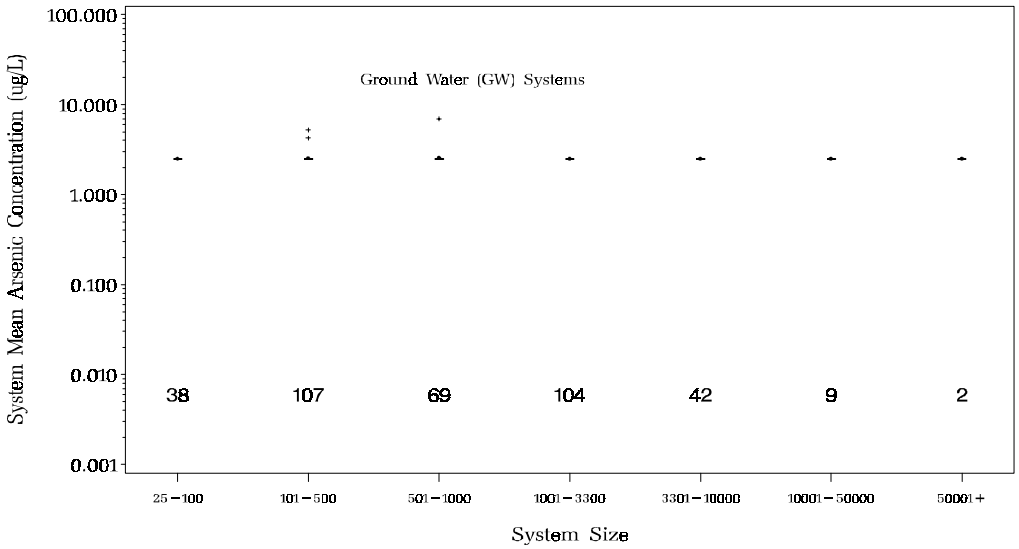


FIGURE 2A. Boxplots of System Means by System Size for
Community Water Systems in State AL
Number of Systems Indicated Below Boxplot



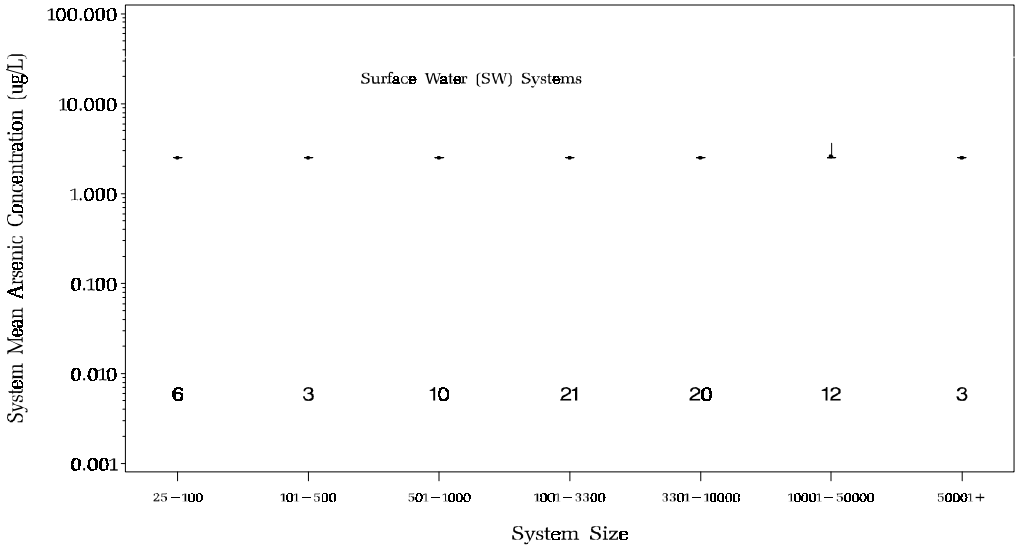
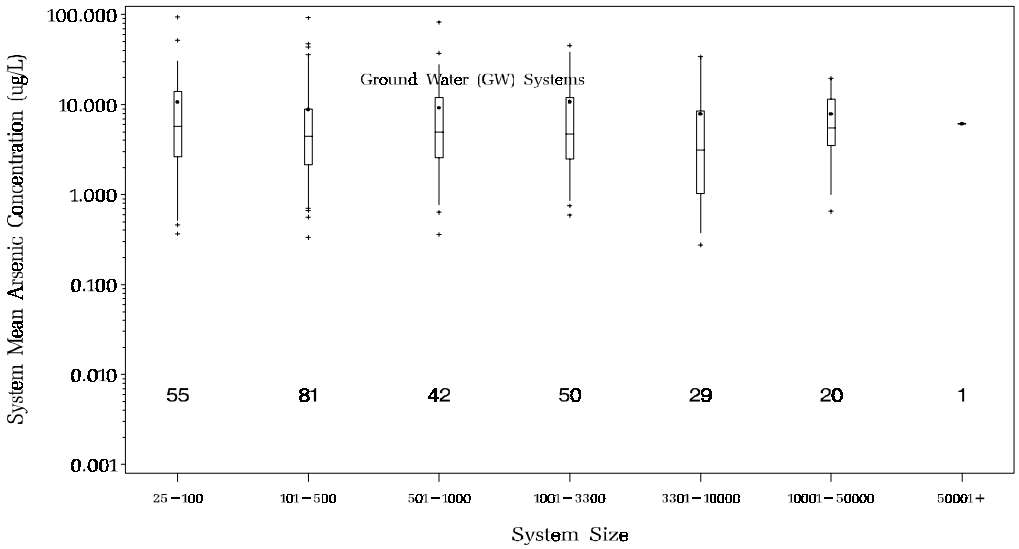


FIGURE 3A. Boxplots of System Means by System Size for
Community Water Systems in State AR
Number of Systems Indicated Below Boxplot



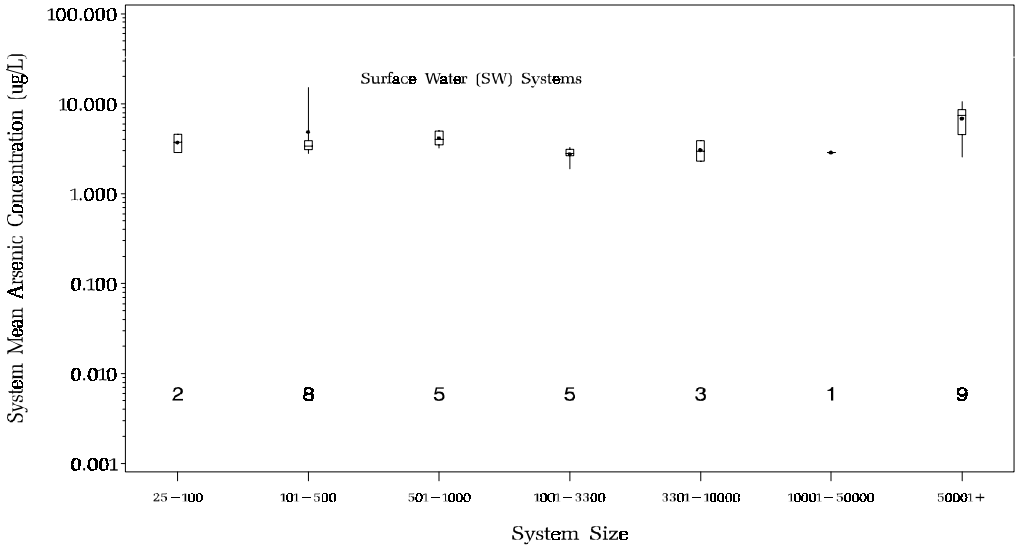
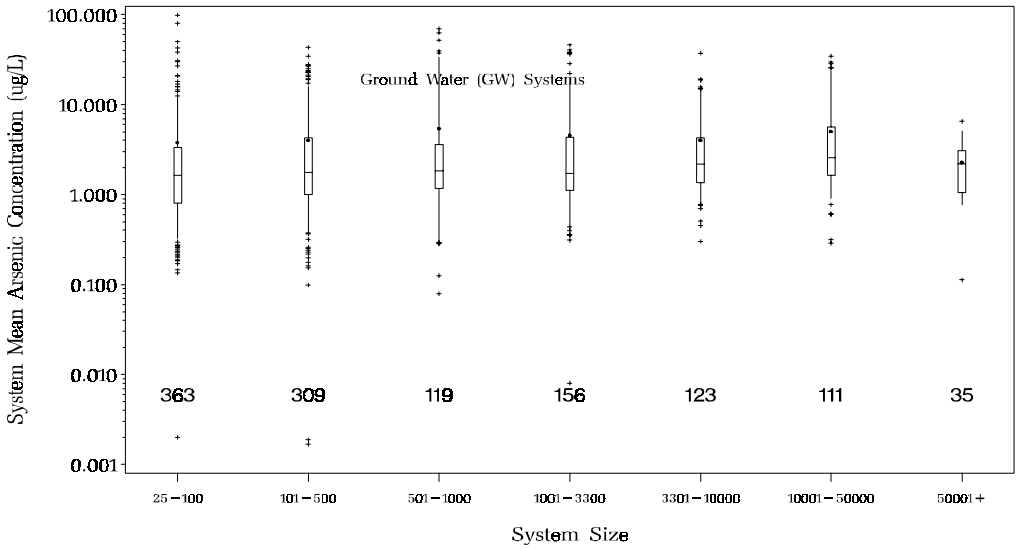


FIGURE 4A. Boxplots of System Means by System Size for
Community Water Systems in State AZ
Number of Systems Indicated Below Boxplot



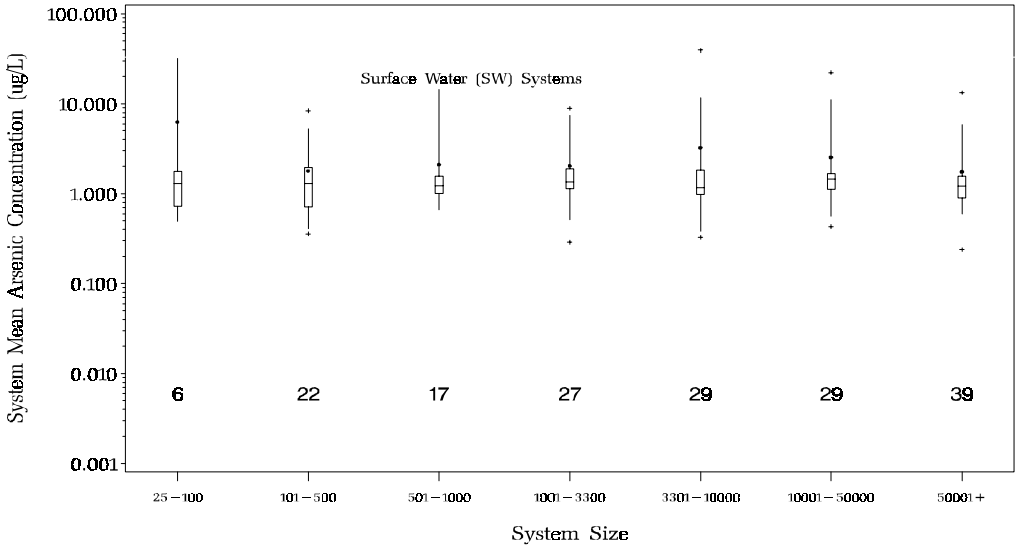


FIGURE 5A. Boxplots of System Means by System Size for
Community Water Systems in State CA
Number of Systems Indicated Below Boxplot

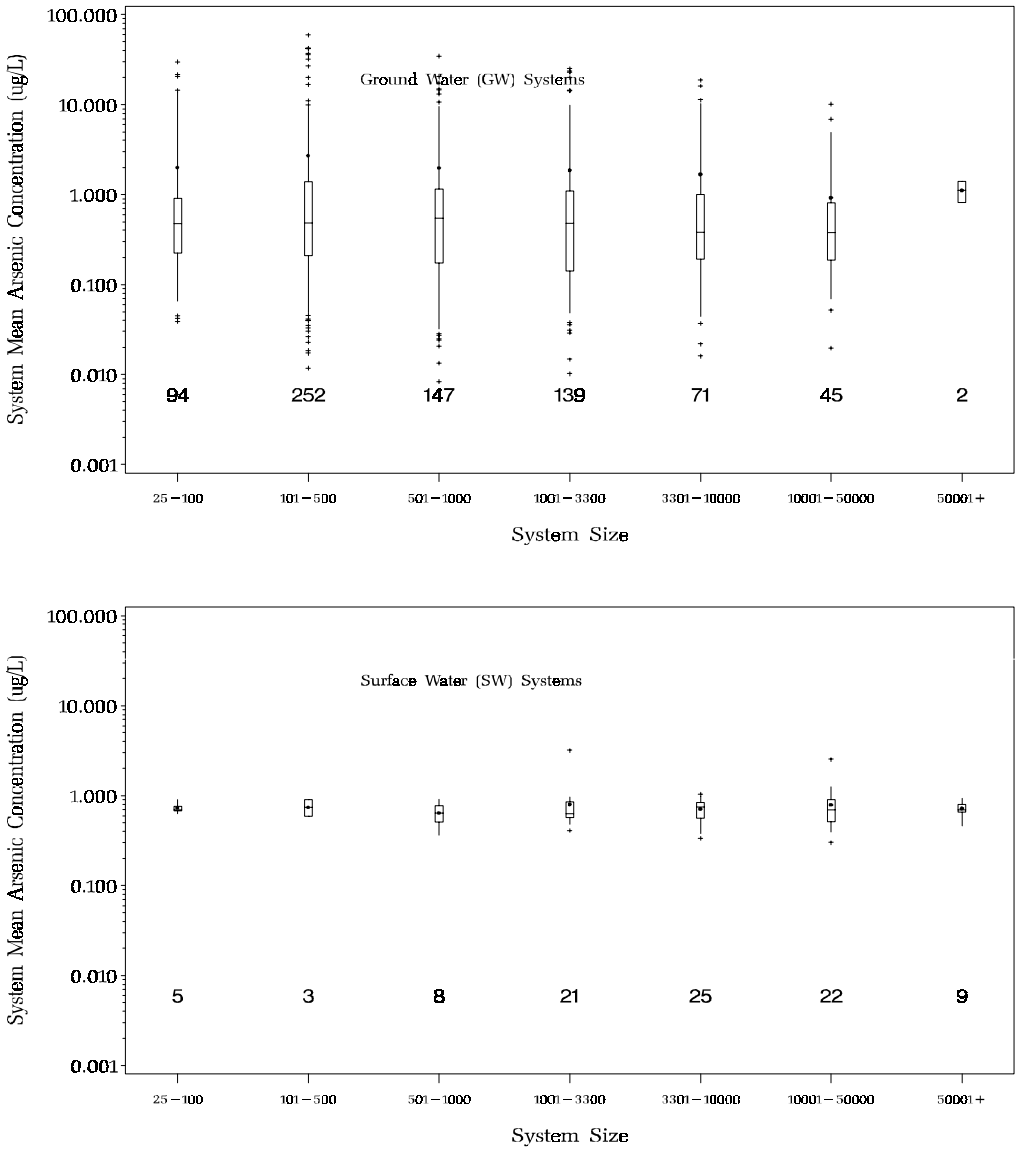
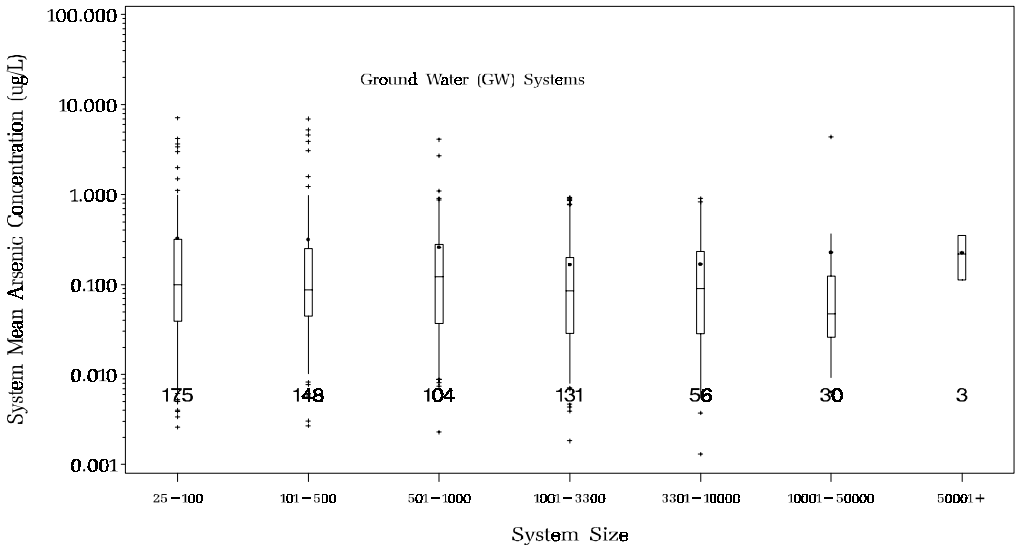


FIGURE 6A. Boxplots of System Means by System Size for
Community Water Systems in State IL
Number of Systems Indicated Below Boxplot



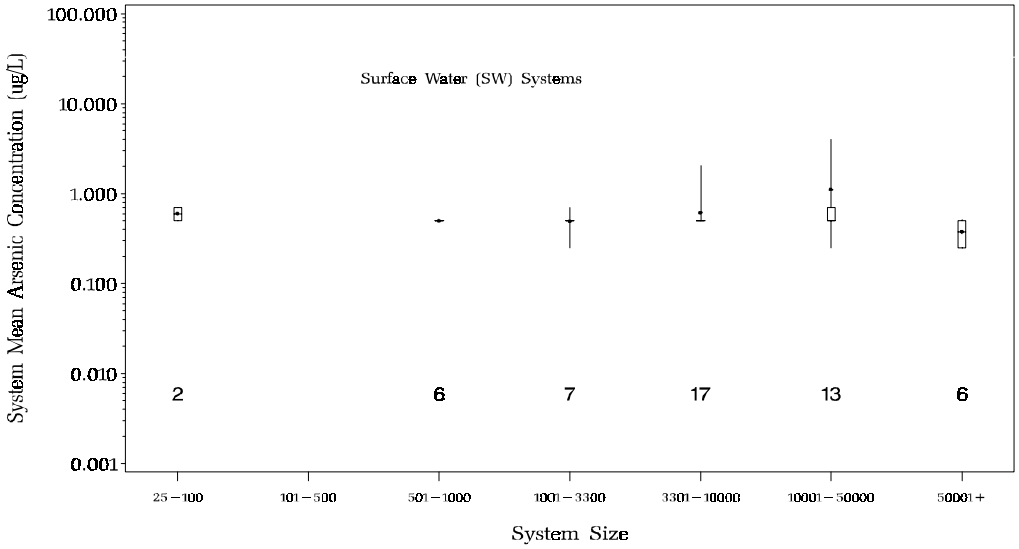
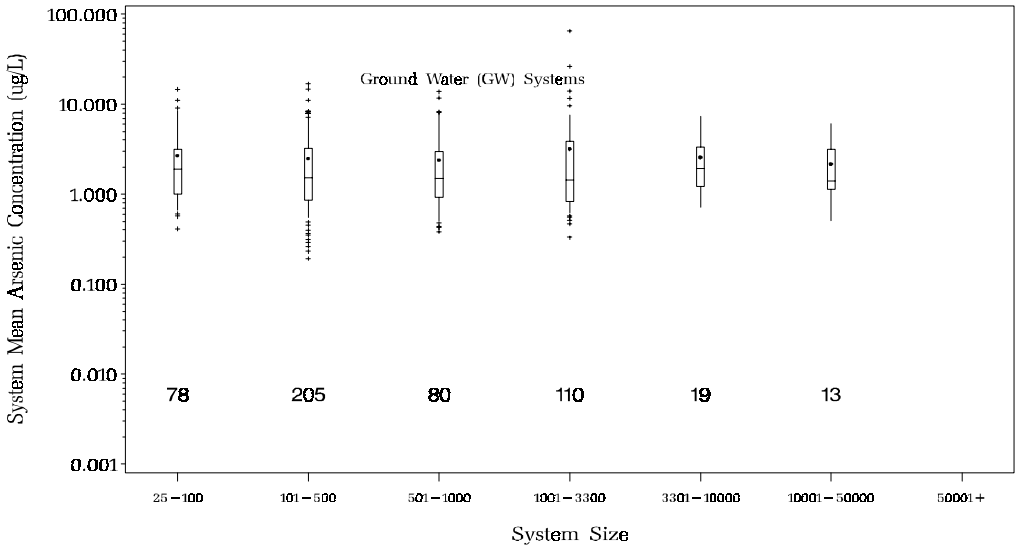


FIGURE 7A. Boxplots of System Means by System Size for
Community Water Systems in State IN
Number of Systems Indicated Below Boxplot



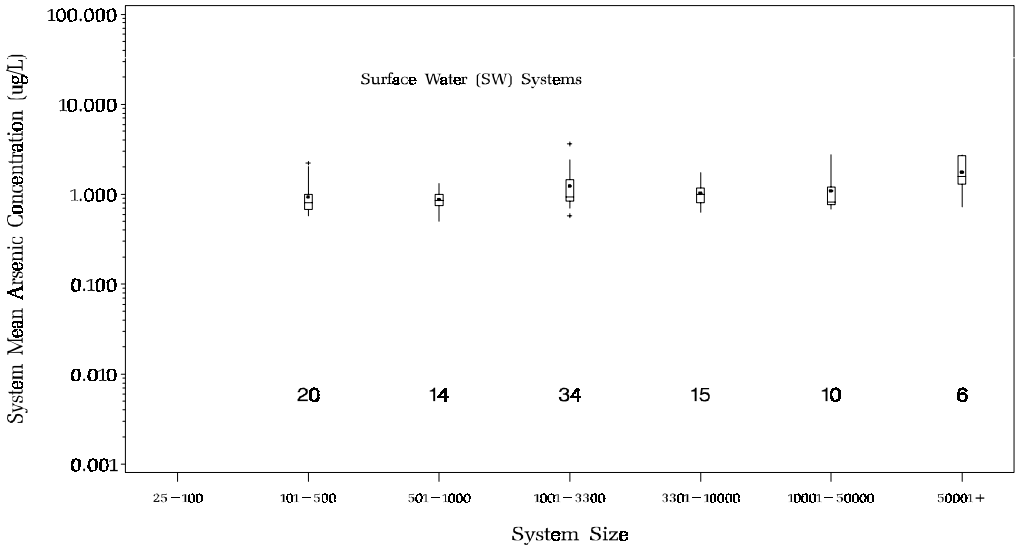
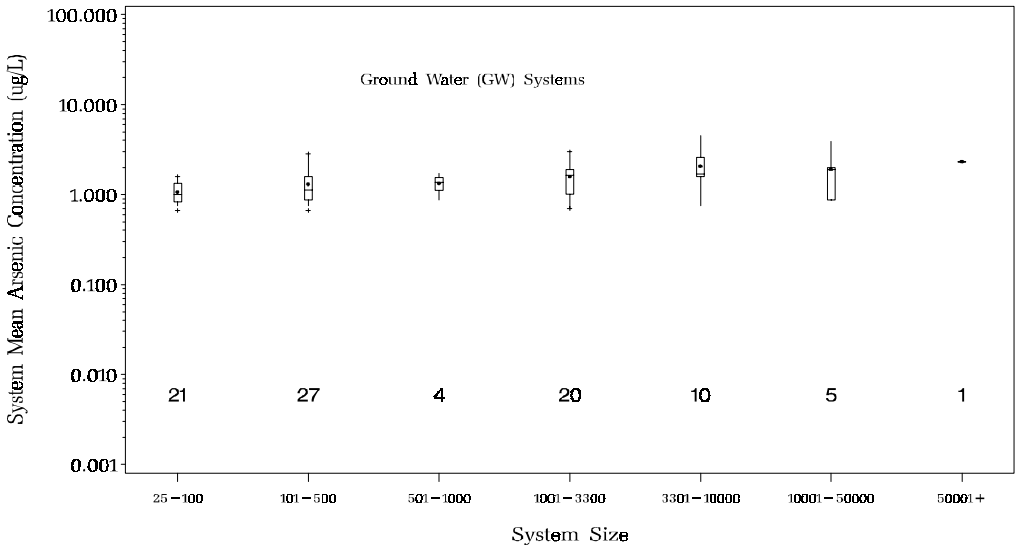


FIGURE 8A. Boxplots of System Means by System Size for
Community Water Systems in State KS
Number of Systems Indicated Below Boxplot



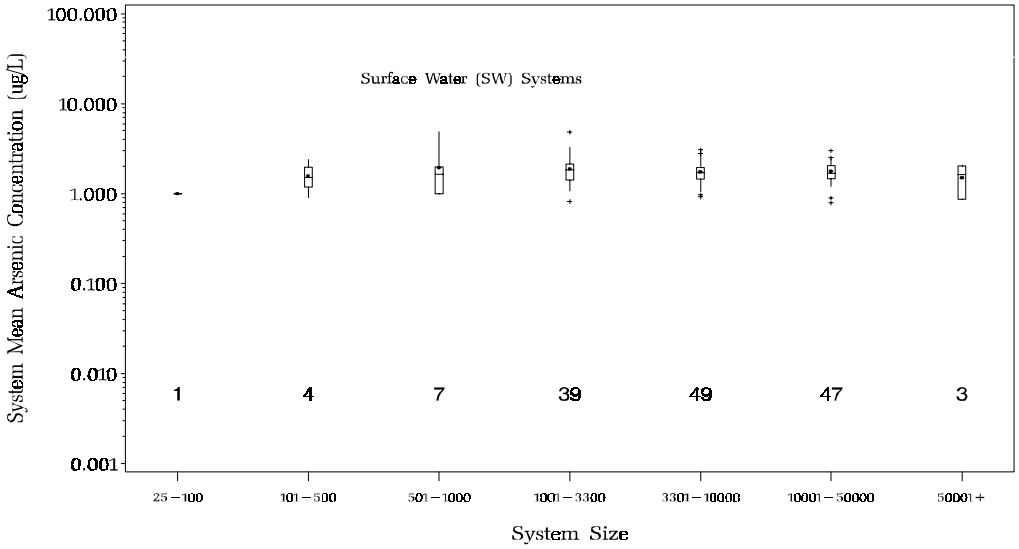
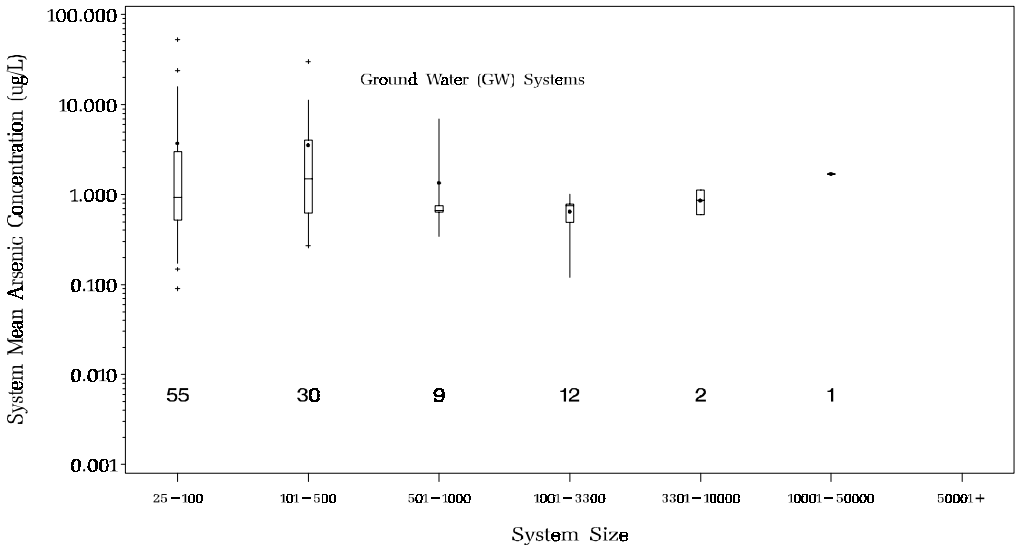


FIGURE 9A. Boxplots of System Means by System Size for
Community Water Systems in State KY
Number of Systems Indicated Below Boxplot



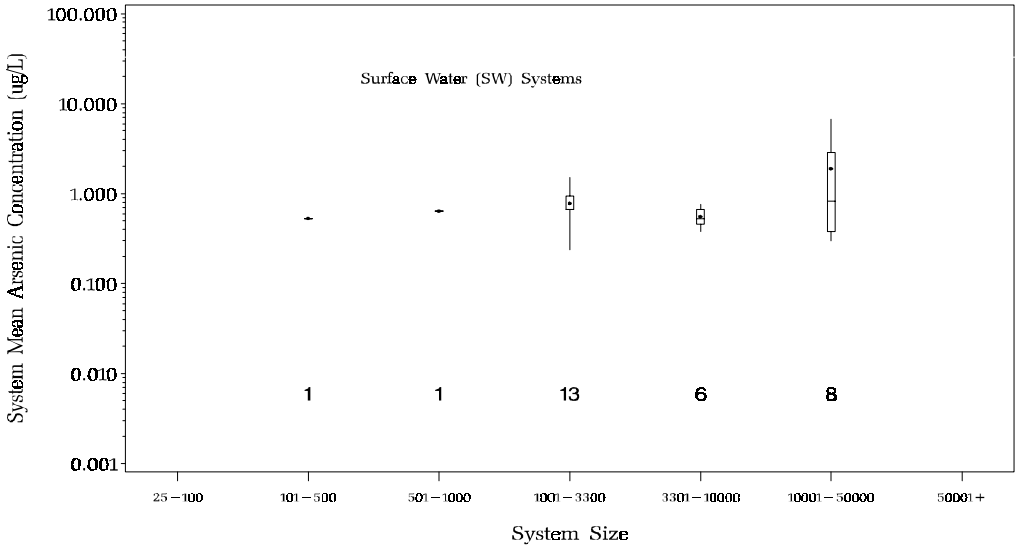
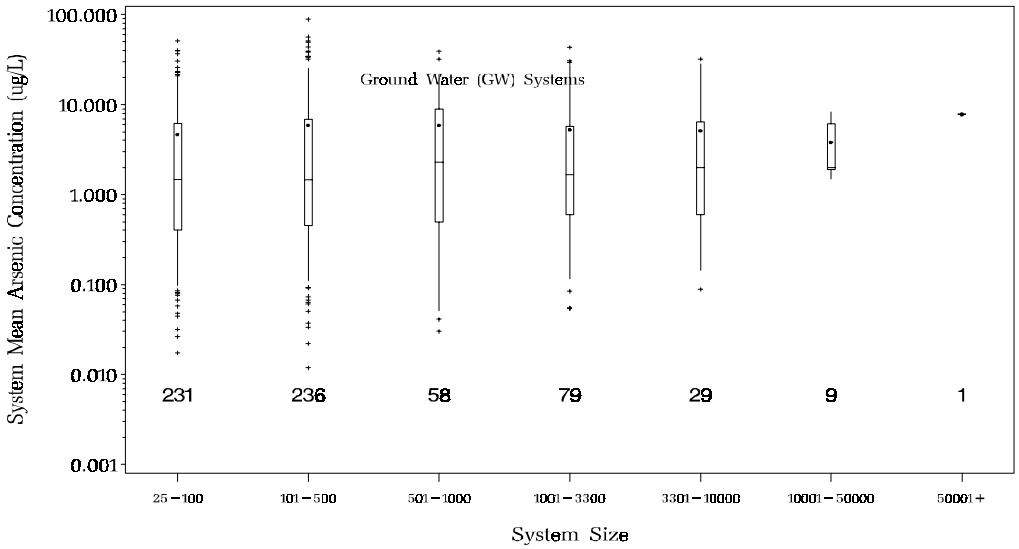


FIGURE 10A. Boxplots of System Means by System Size for
Community Water Systems in State ME
Number of Systems Indicated Below Boxplot



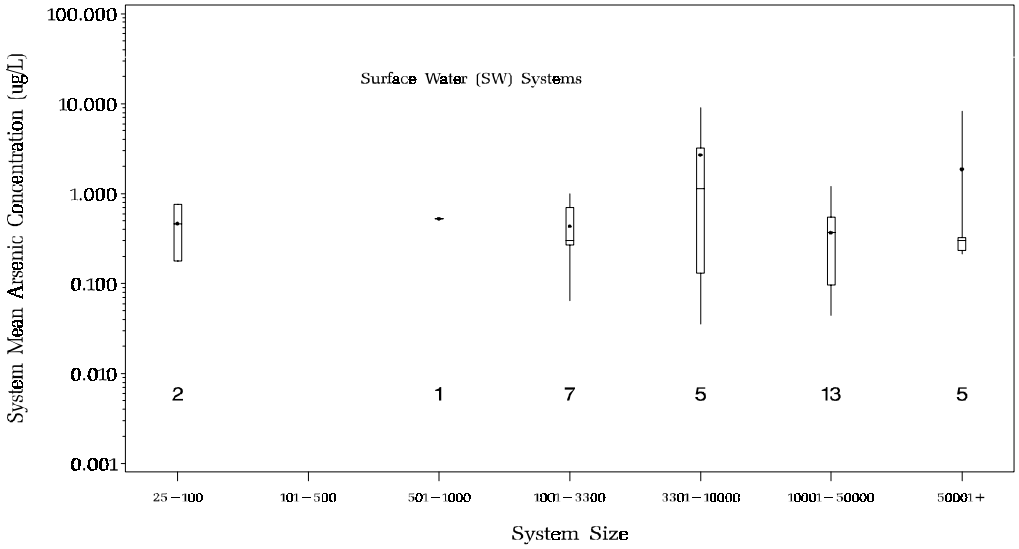
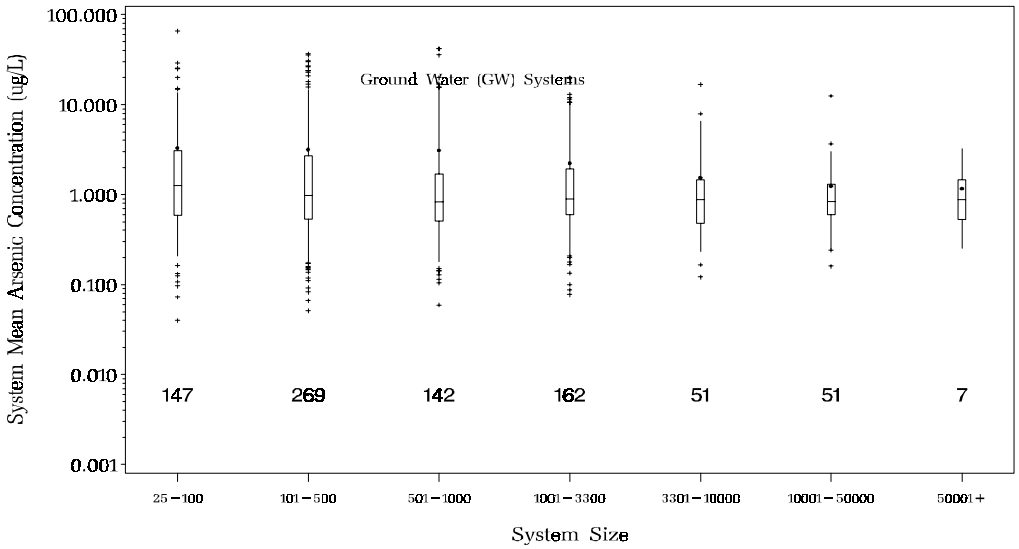


FIGURE 11A. Boxplots of System Means by System Size for
Community Water Systems in State MI
Number of Systems Indicated Below Boxplot



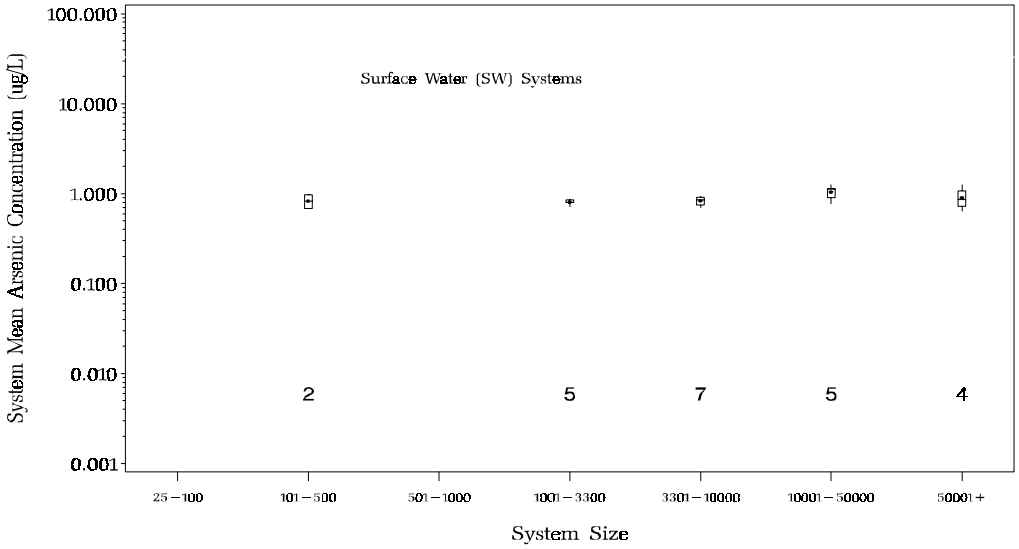
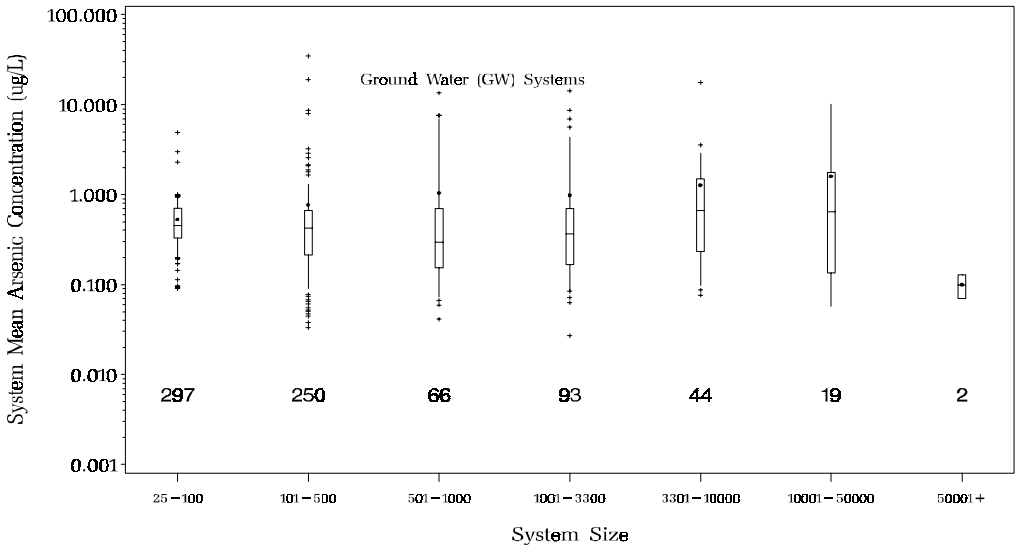


FIGURE 12A. Boxplots of System Means by System Size for
Community Water Systems in State MN
Number of Systems Indicated Below Boxplot



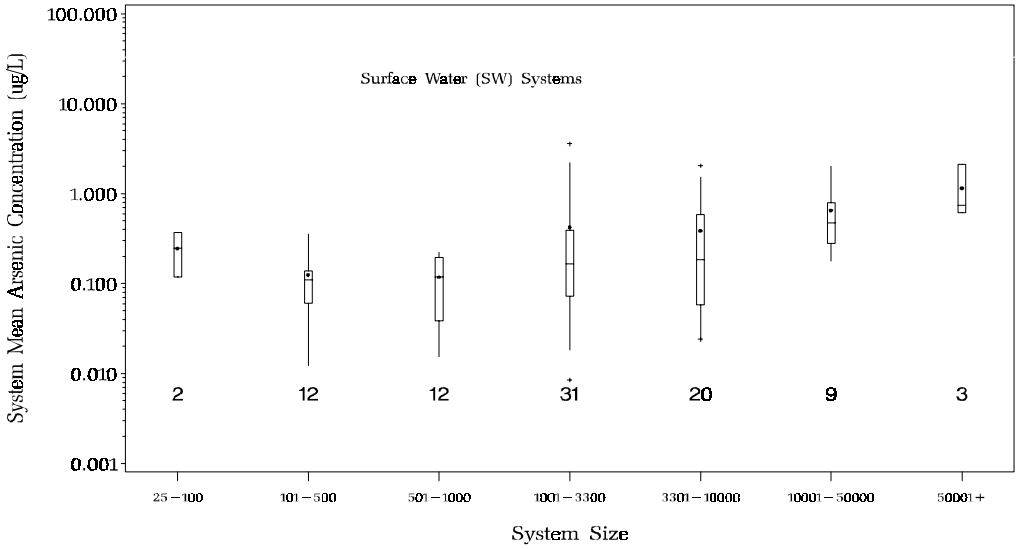
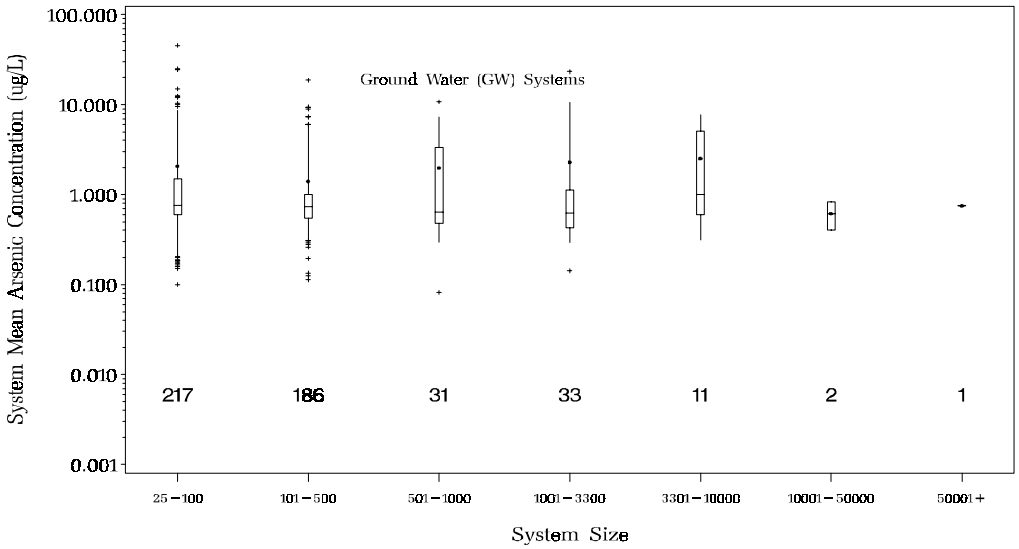


FIGURE 13A. Boxplots of System Means by System Size for
Community Water Systems in State MO
Number of Systems Indicated Below Boxplot



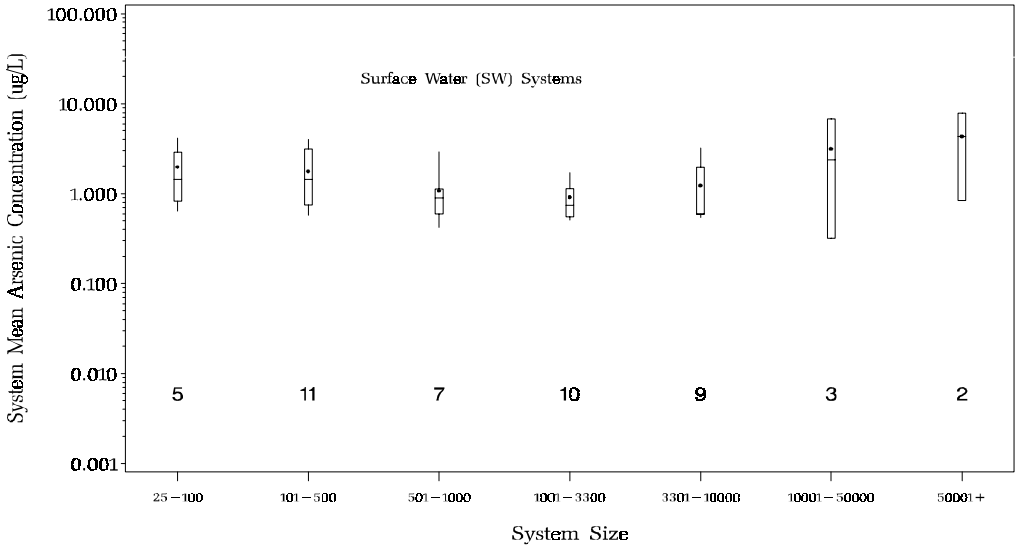
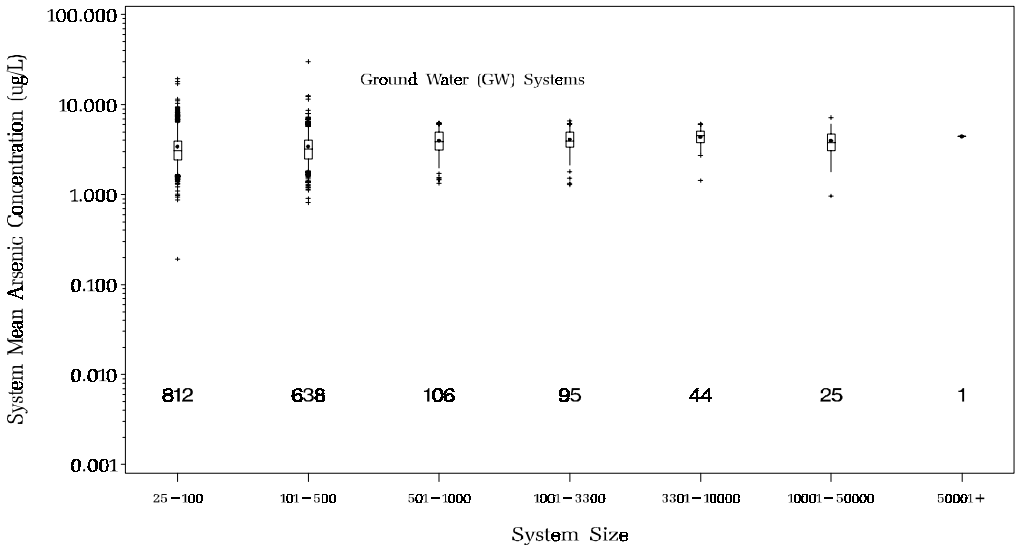


FIGURE 14A. Boxplots of System Means by System Size for
Community Water Systems in State MT
Number of Systems Indicated Below Boxplot



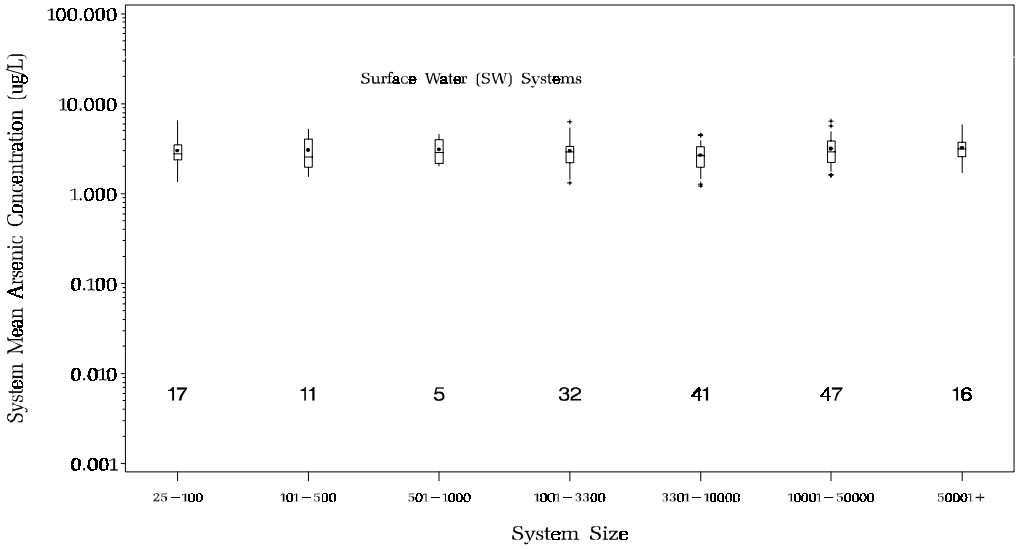
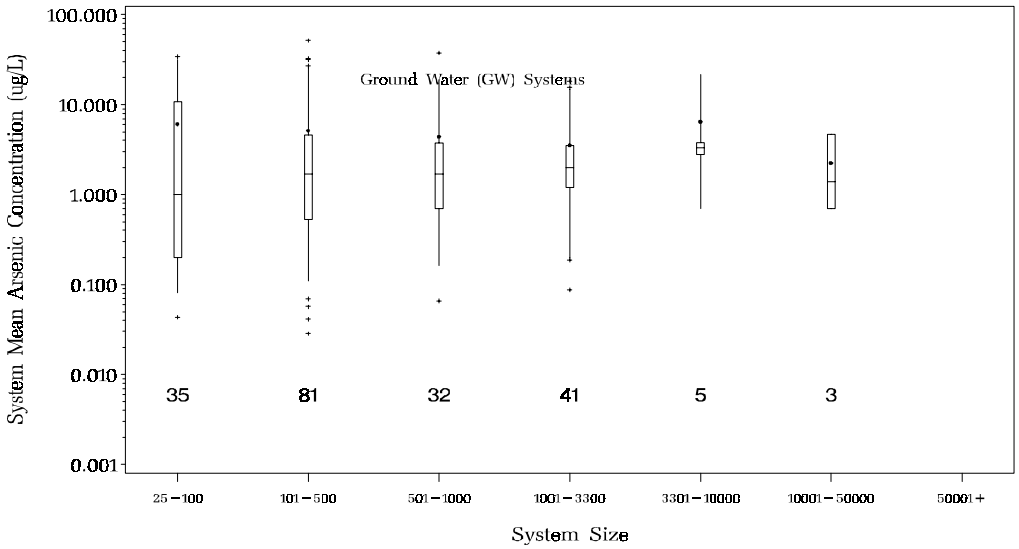


FIGURE 15A. Boxplots of System Means by System Size for
Community Water Systems in State NC
Number of Systems Indicated Below Boxplot



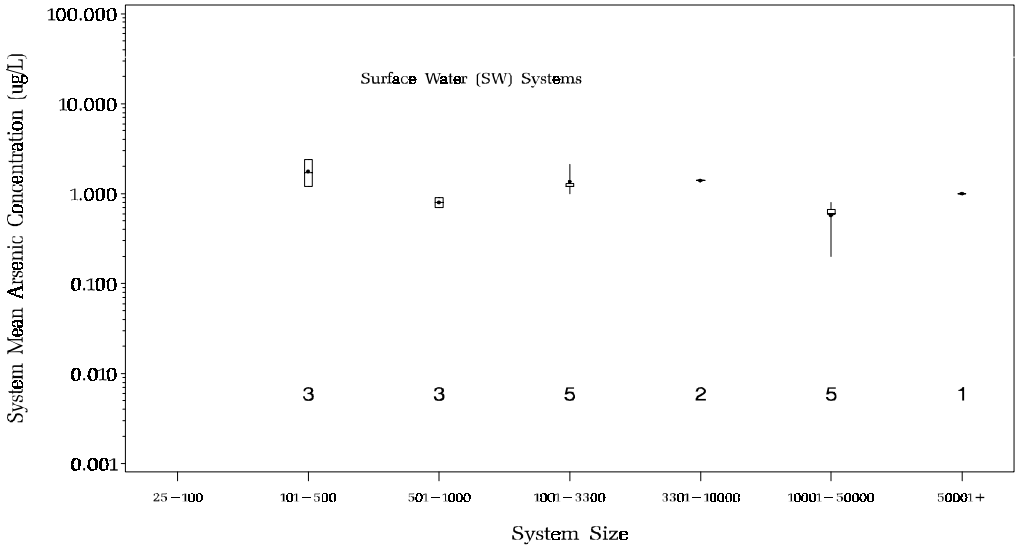
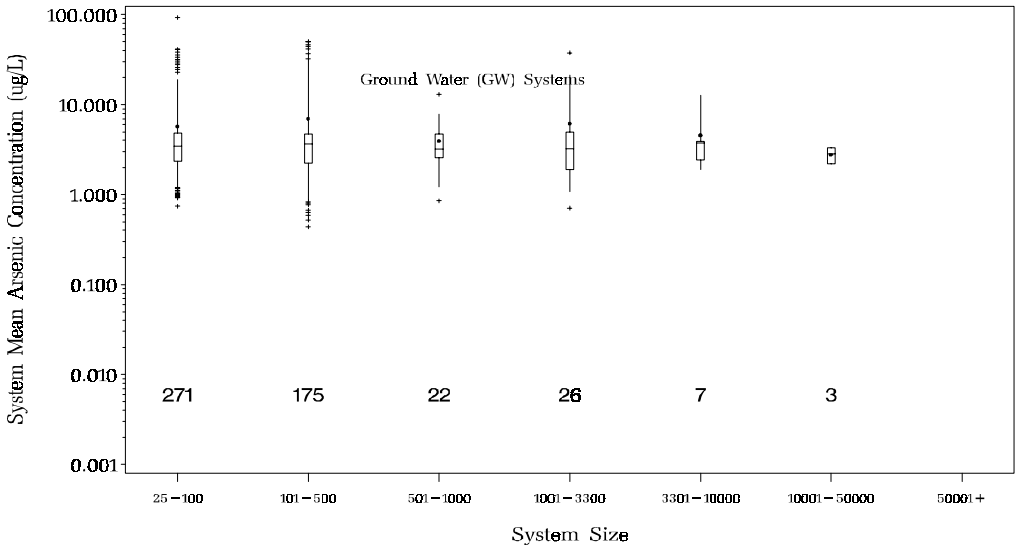


FIGURE 16A. Boxplots of System Means by System Size for
Community Water Systems in State ND
Number of Systems Indicated Below Boxplot



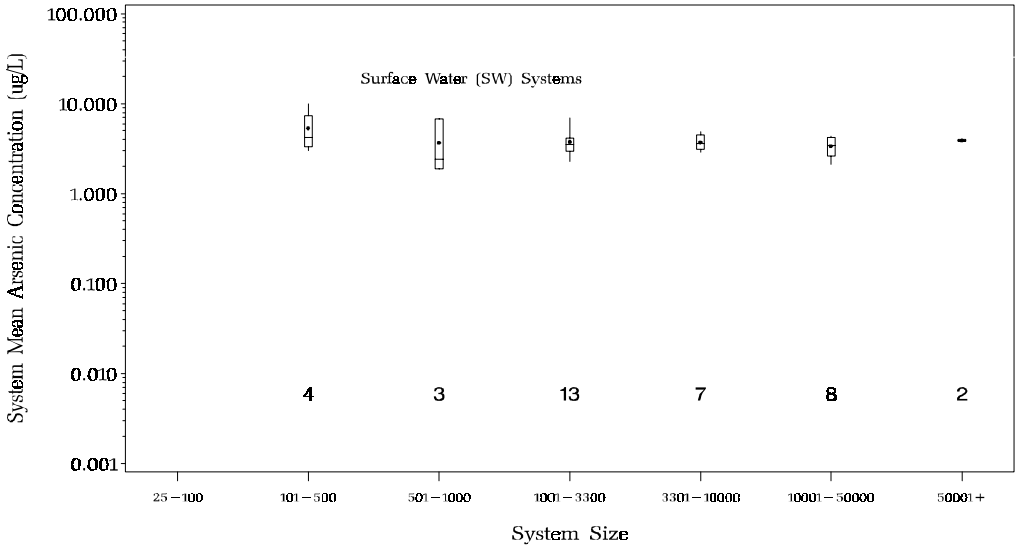
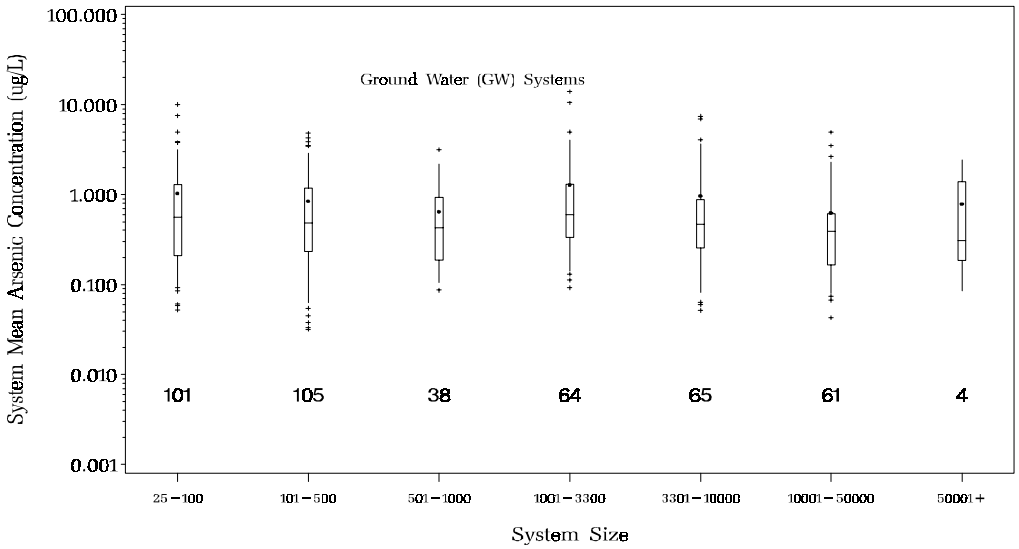


FIGURE 17A. Boxplots of System Means by System Size for
Community Water Systems in State NH
Number of Systems Indicated Below Boxplot



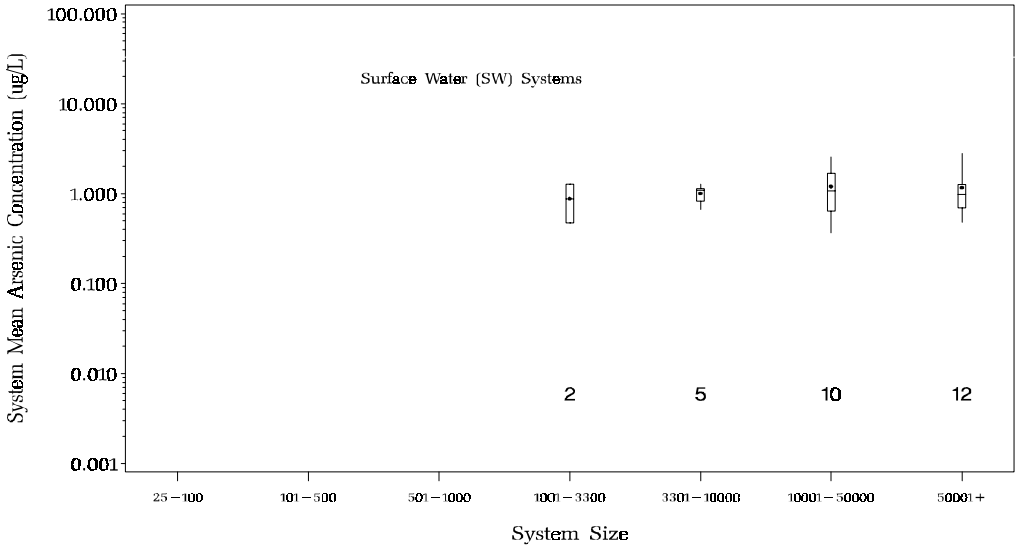
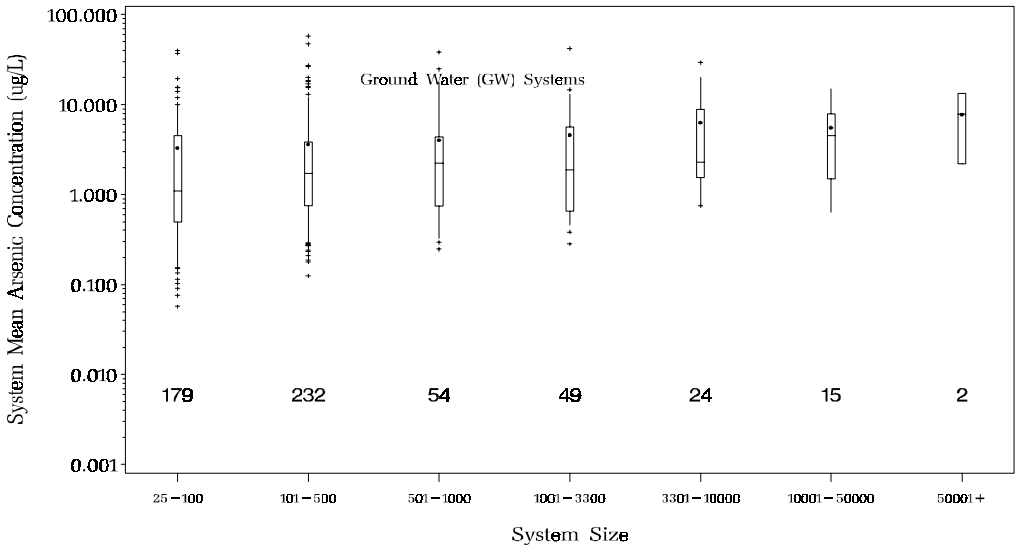


FIGURE 18A. Boxplots of System Means by System Size for
Community Water Systems in State NJ
Number of Systems Indicated Below Boxplot



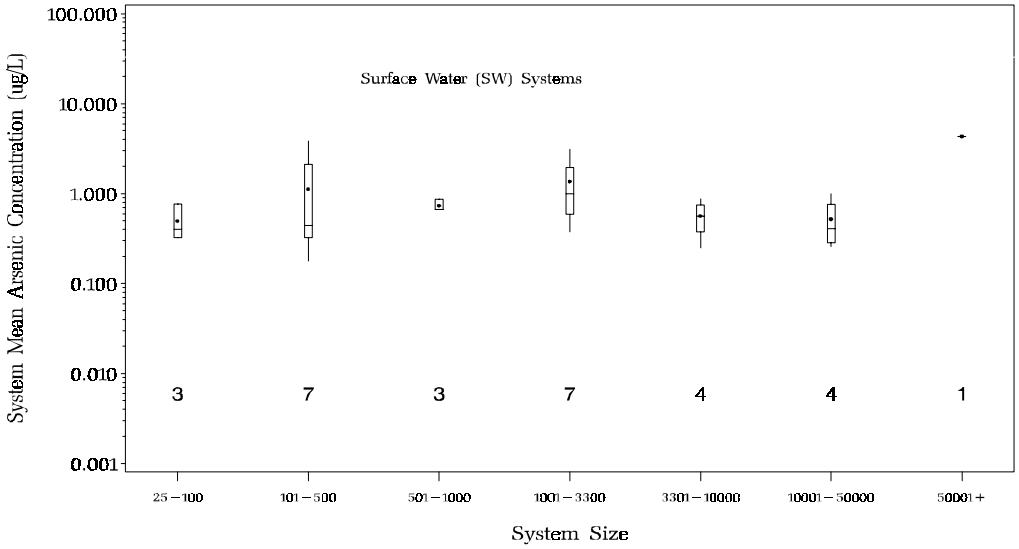
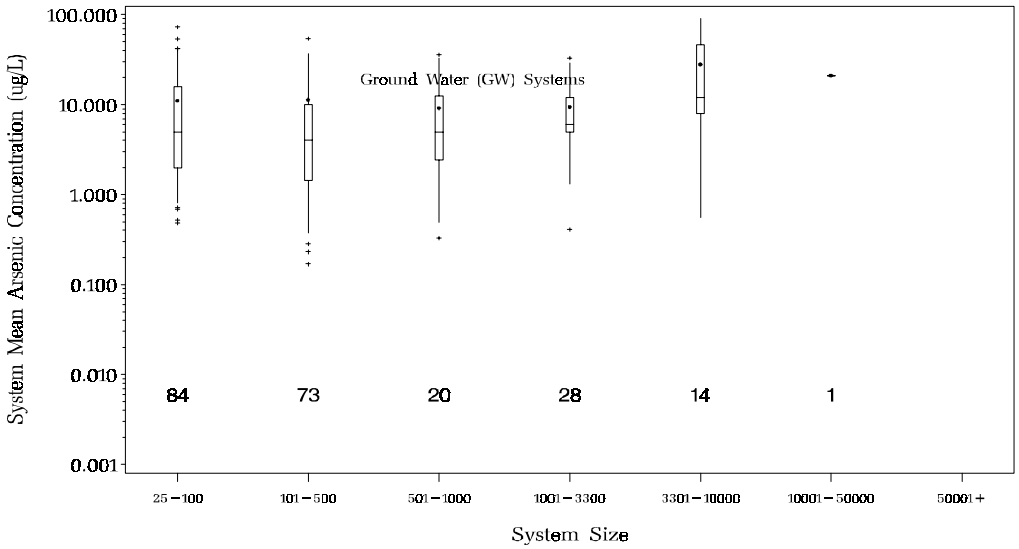


FIGURE 19A. Boxplots of System Means by System Size for
Community Water Systems in State NM
Number of Systems Indicated Below Boxplot



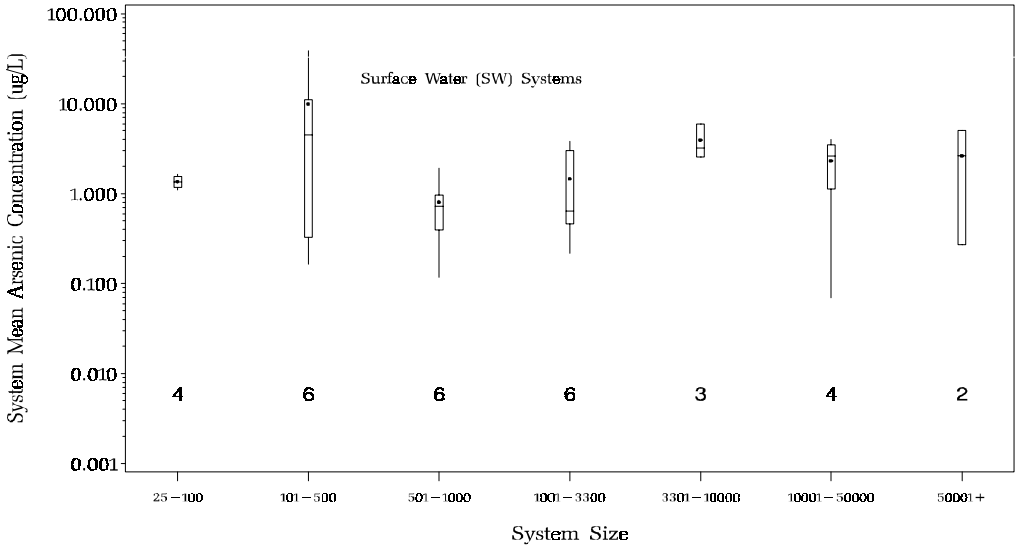


FIGURE 20A. Boxplots of System Means by System Size for
Community Water Systems in State NV
Number of Systems Indicated Below Boxplot

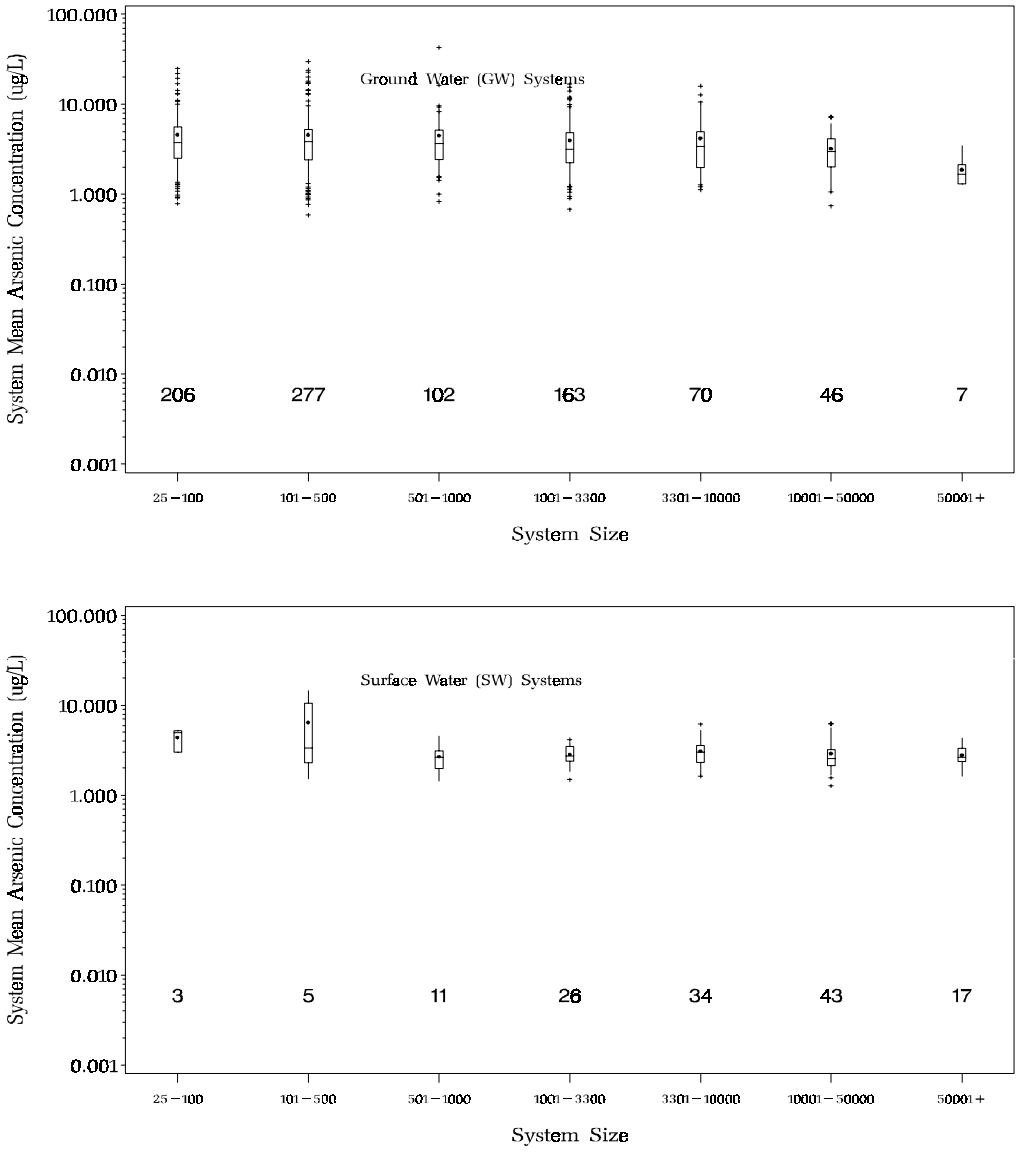
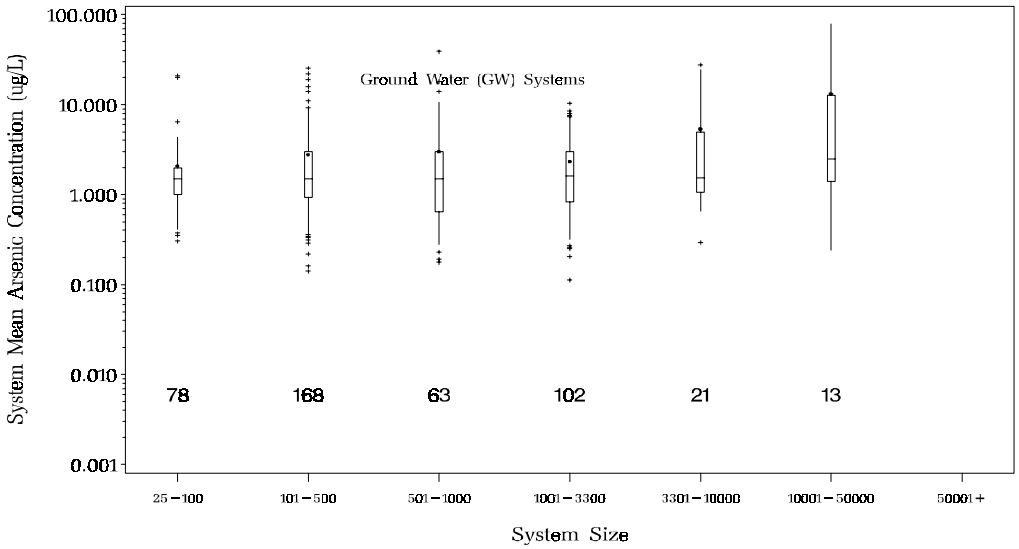


FIGURE 21A. Boxplots of System Means by System Size for Community Water Systems in State OH
Number of Systems Indicated Below Boxplot



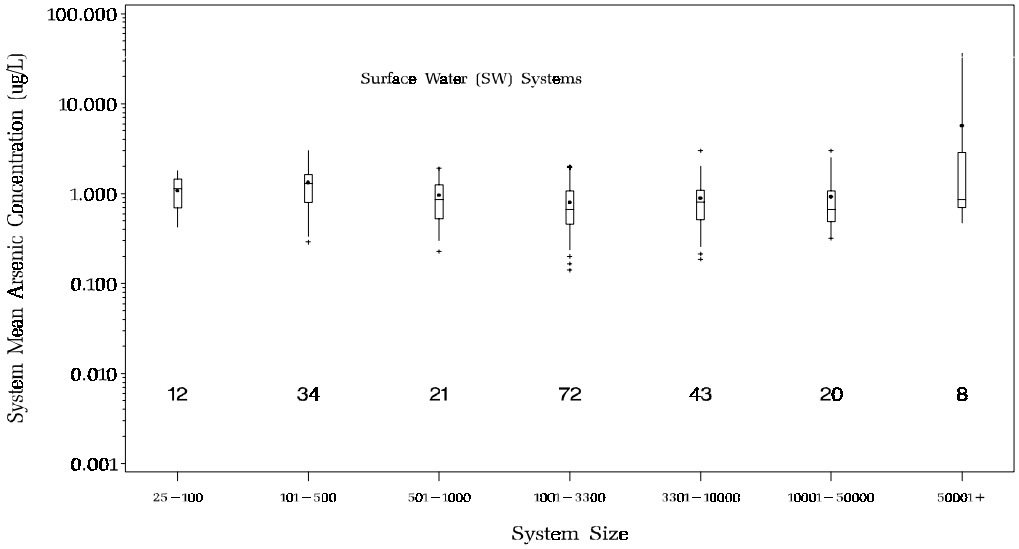
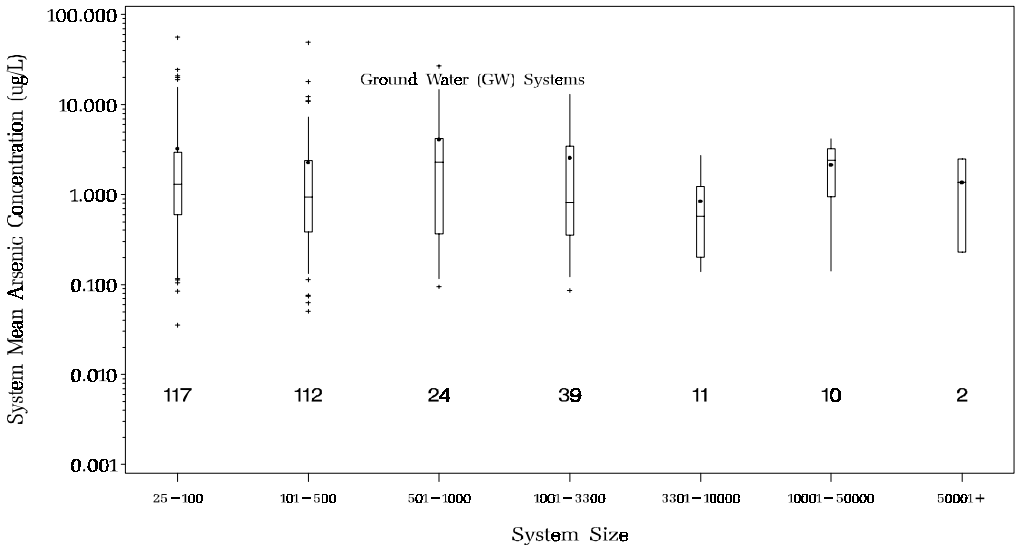


FIGURE 22A. Boxplots of System Means by System Size for
Community Water Systems in State OK
Number of Systems Indicated Below Boxplot



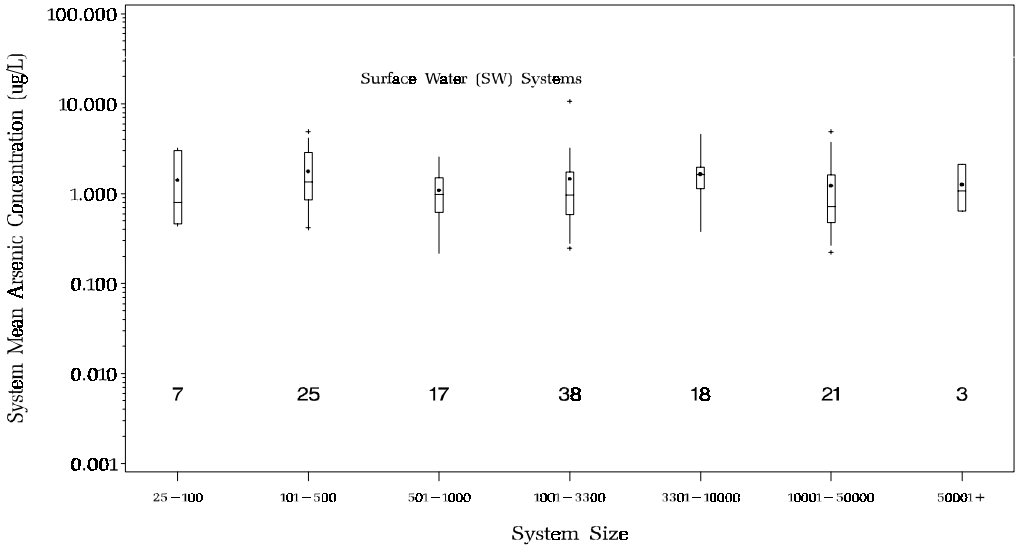
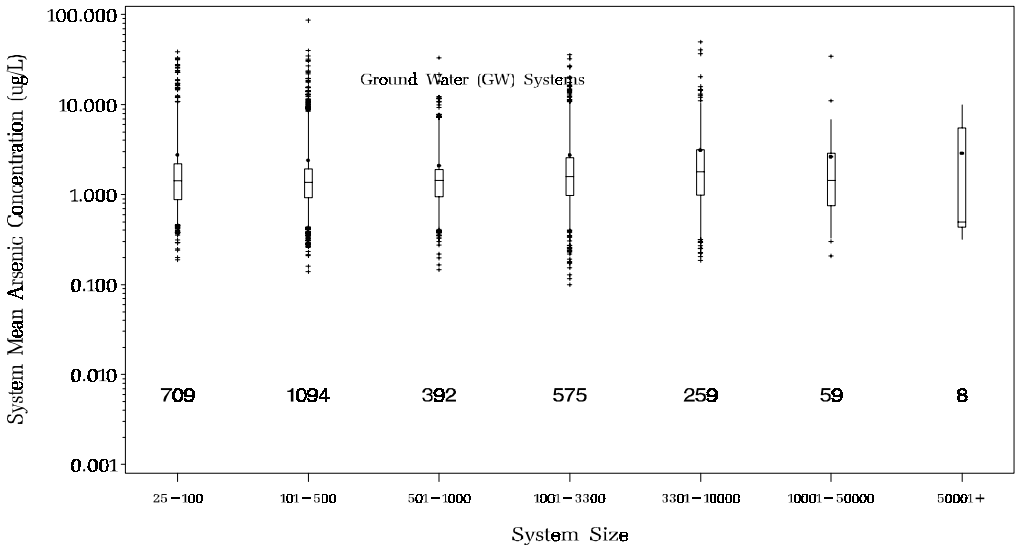


FIGURE 23A. Boxplots of System Means by System Size for
Community Water Systems in State OR
Number of Systems Indicated Below Boxplot



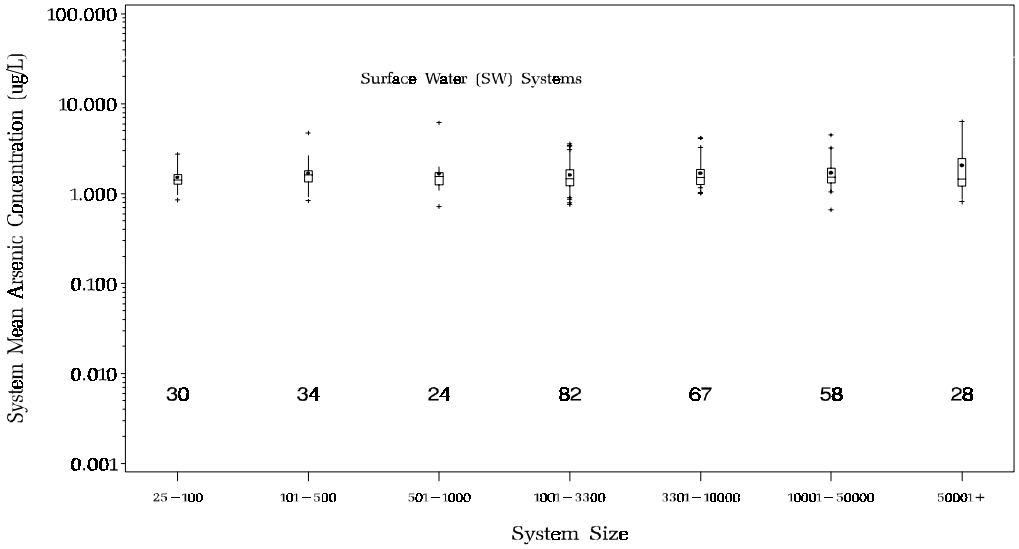
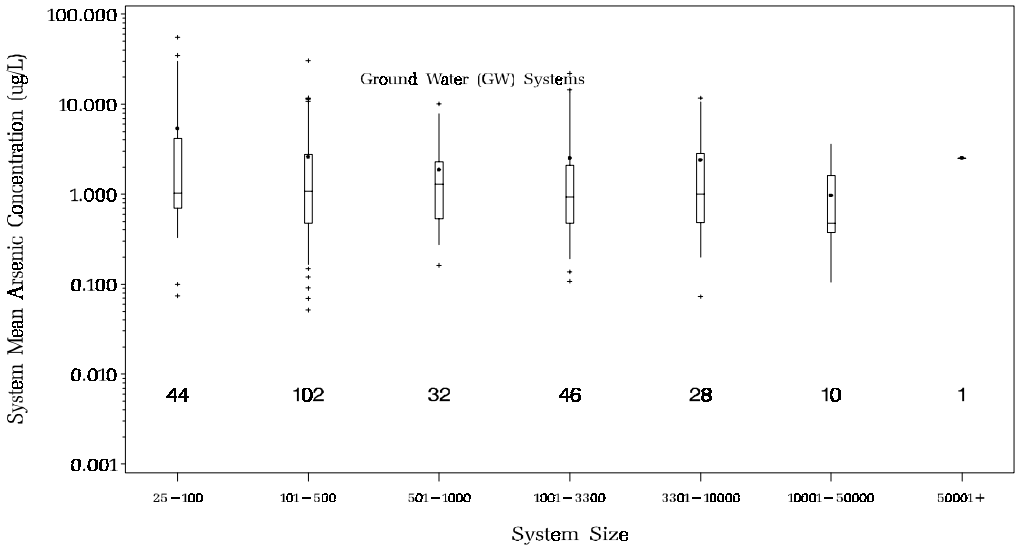


FIGURE 24A. Boxplots of System Means by System Size for
Community Water Systems in State TX
Number of Systems Indicated Below Boxplot



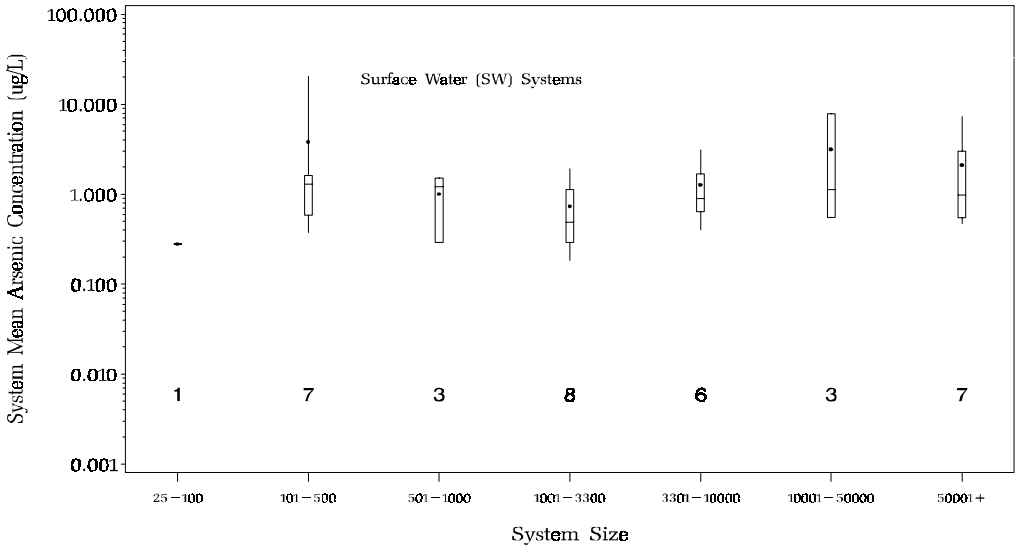


FIGURE 25A. Boxplots of System Means by System Size for
Community Water Systems in State UT
Number of Systems Indicated Below Boxplot

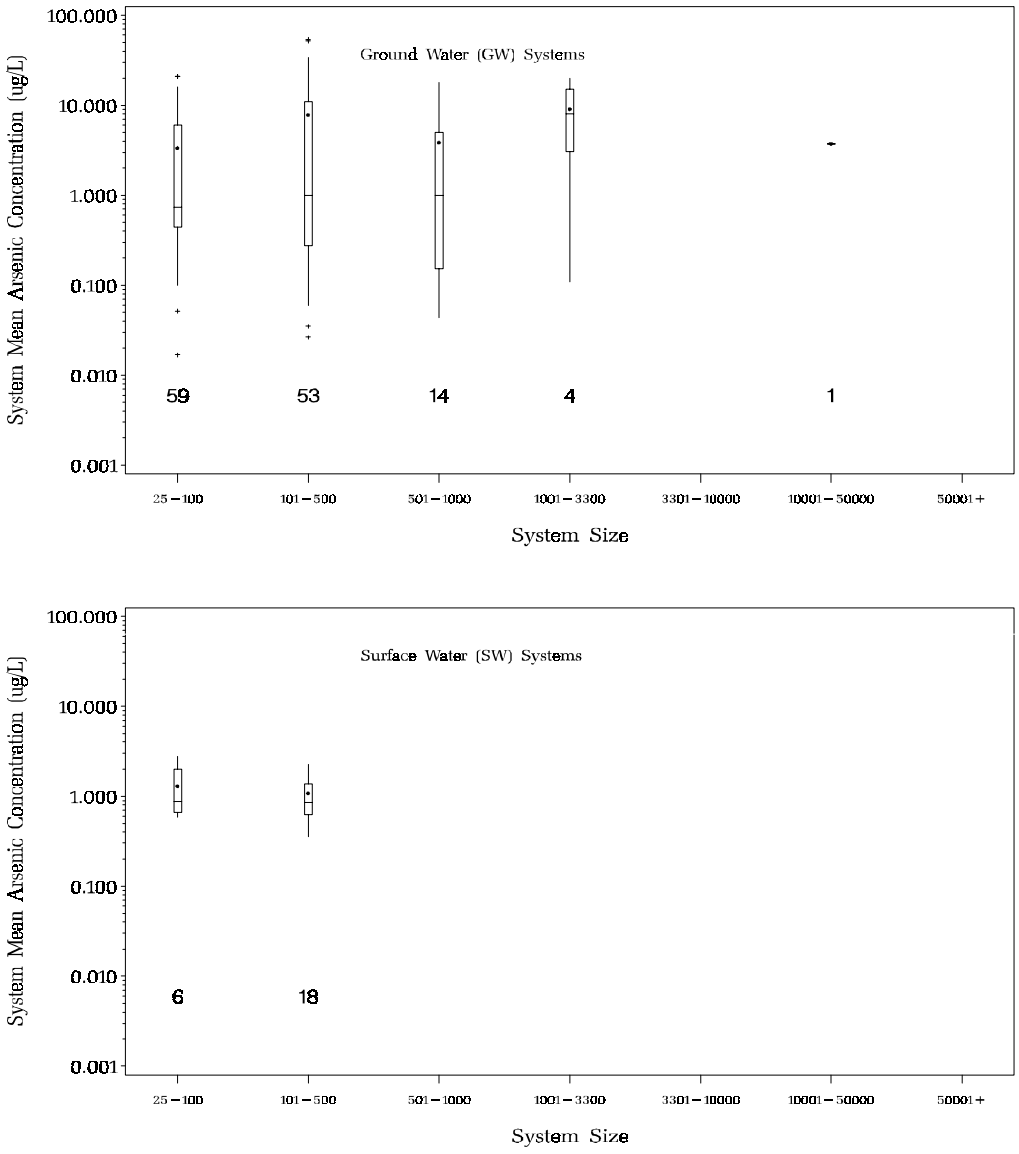


FIGURE 1B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State AK
Number of Systems Indicated Below Boxplot

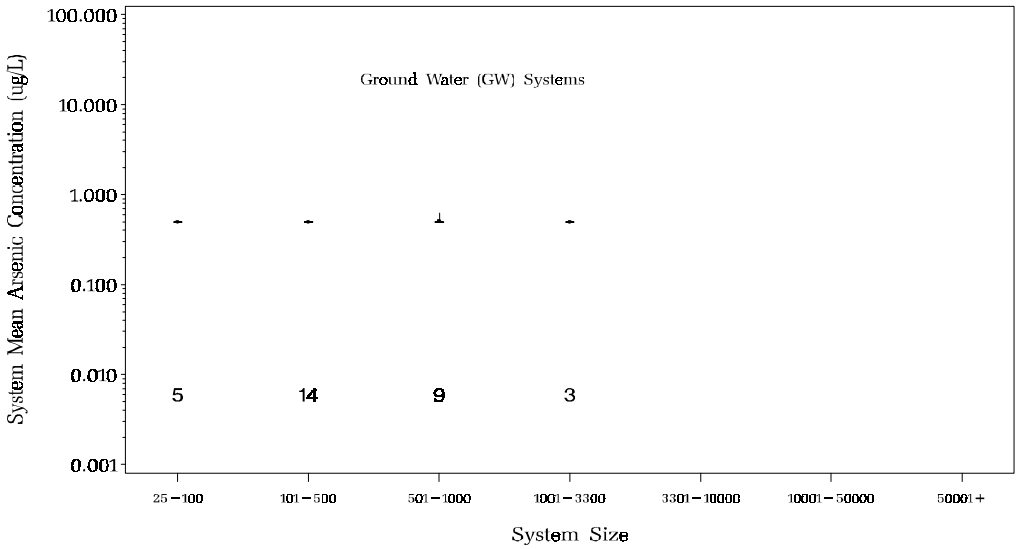
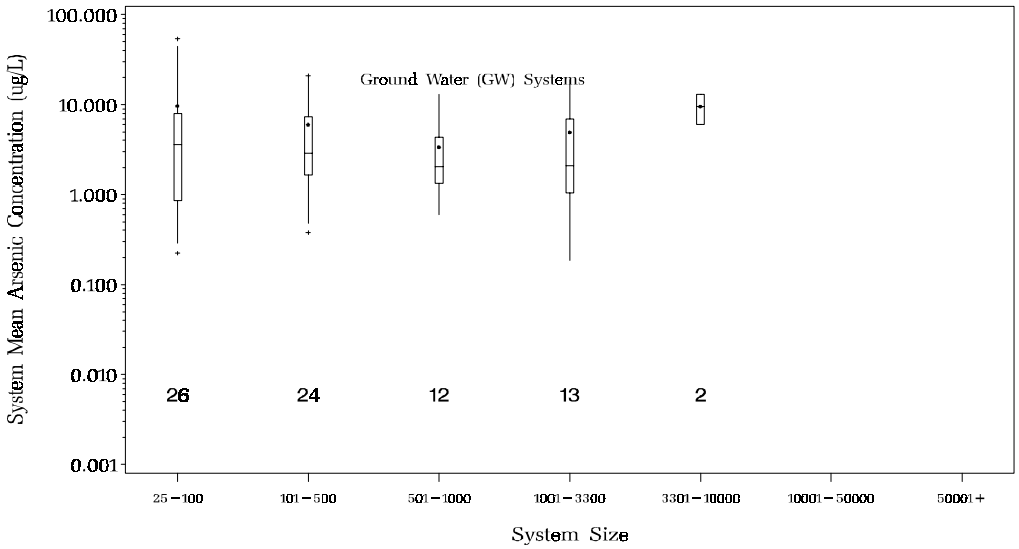


FIGURE 2B. Boxplots of System Means by System Size for Non – Transient Non – Community Water Systems in State AL. Number of Systems Indicated Below Boxplot



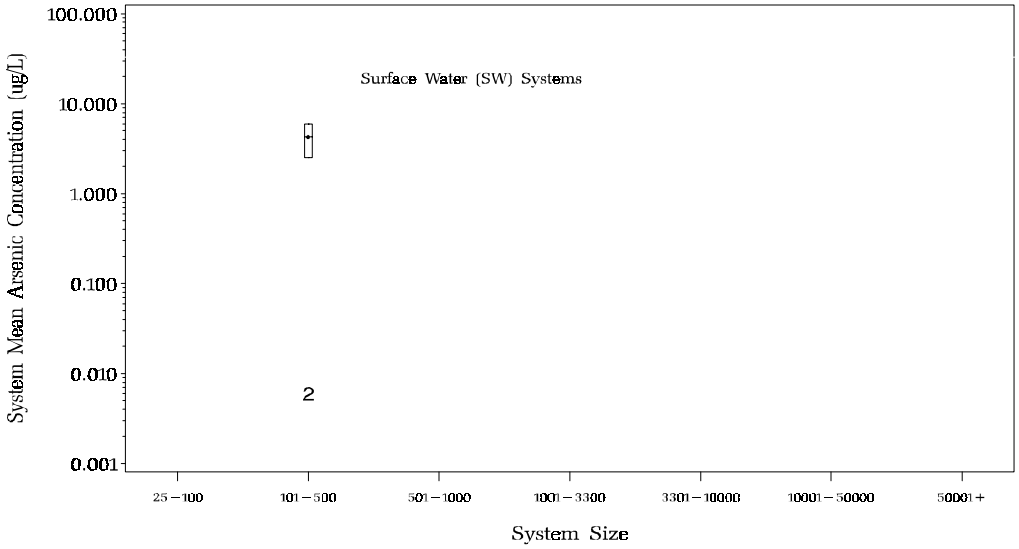


FIGURE 3B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State AZ
Number of Systems Indicated Below Boxplot

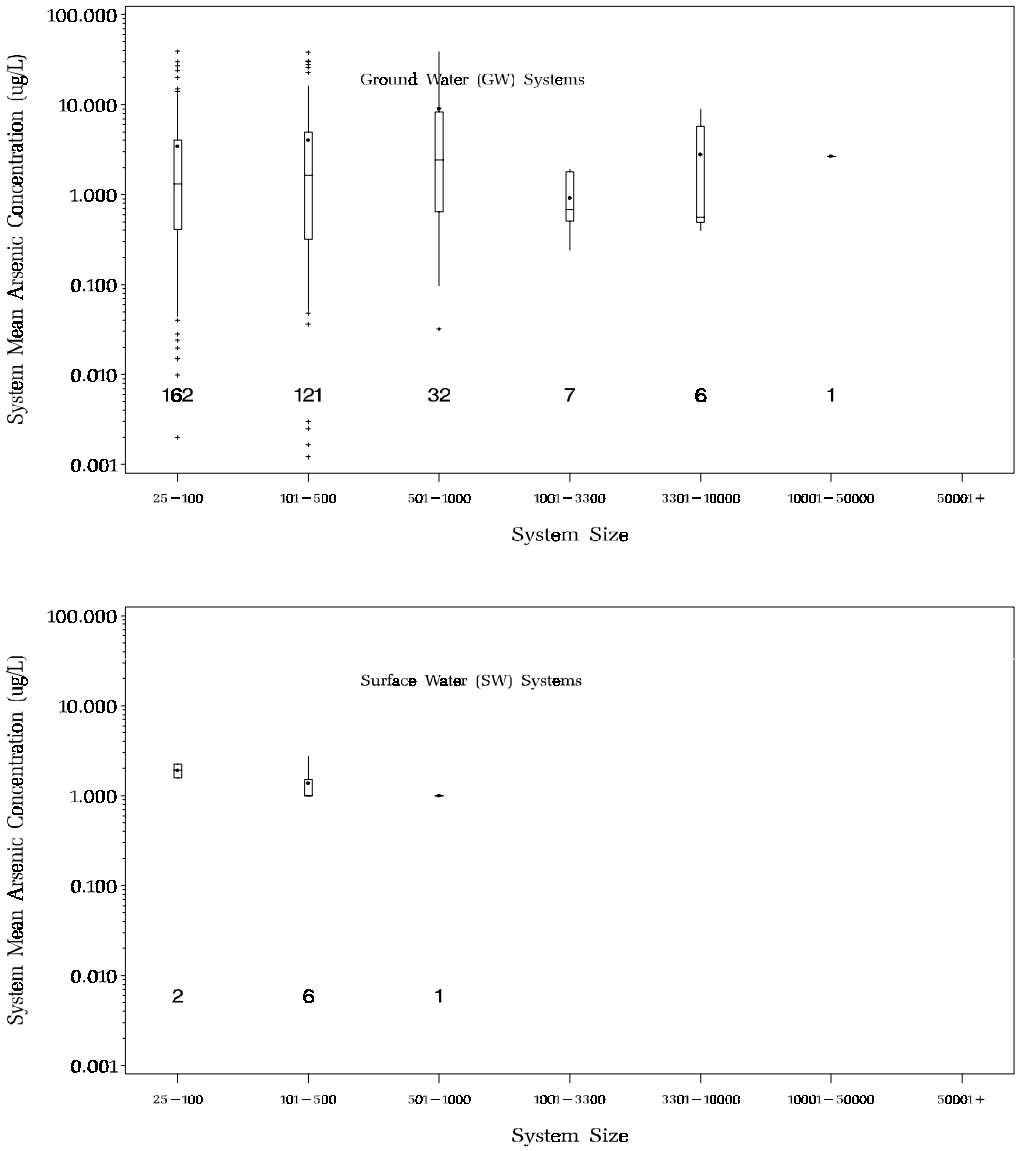


FIGURE 4B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State CA
Number of Systems Indicated Below Boxplot

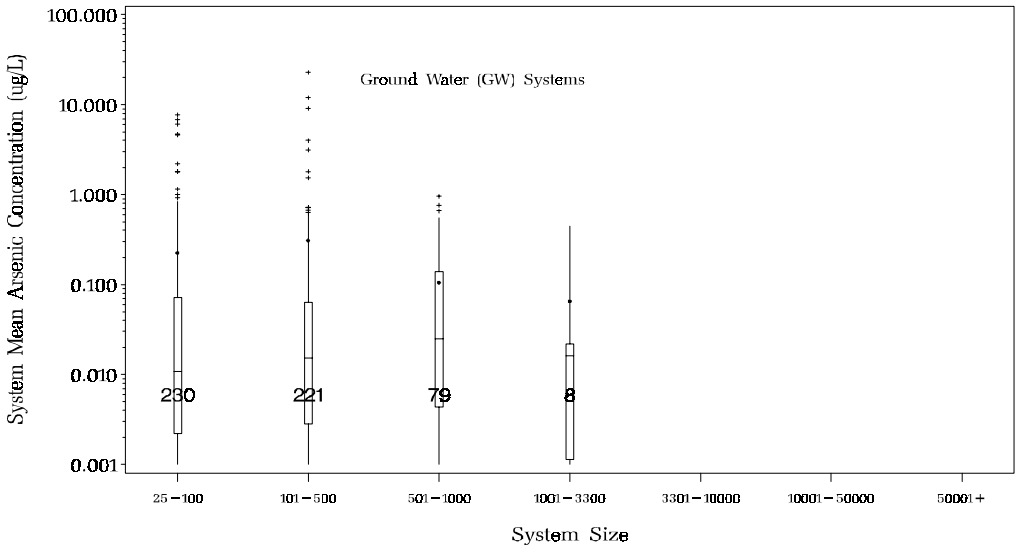
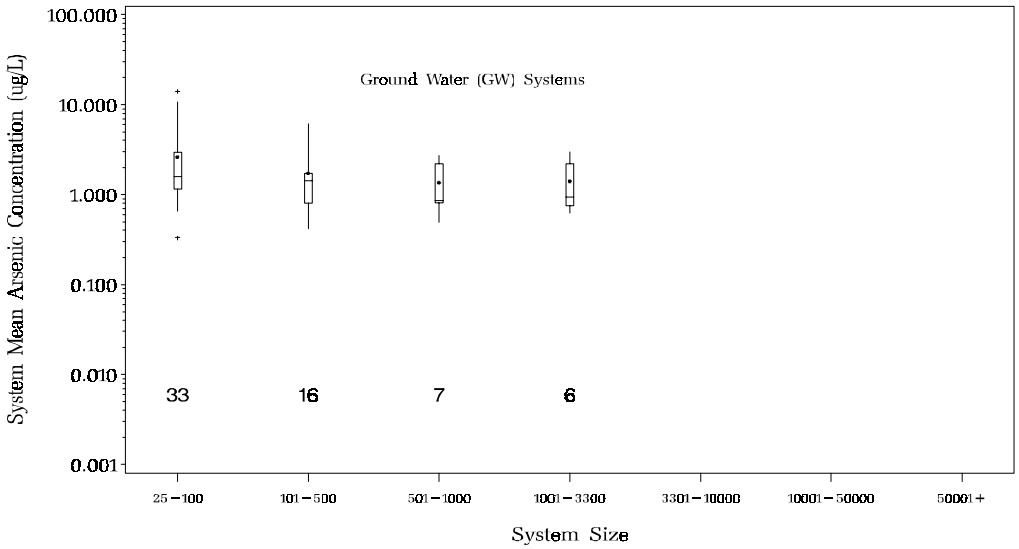


FIGURE 5B. Boxplots of System Means by System Size for Non–Transient Non–Community Water Systems in State IN Number of Systems Indicated Below Boxplot



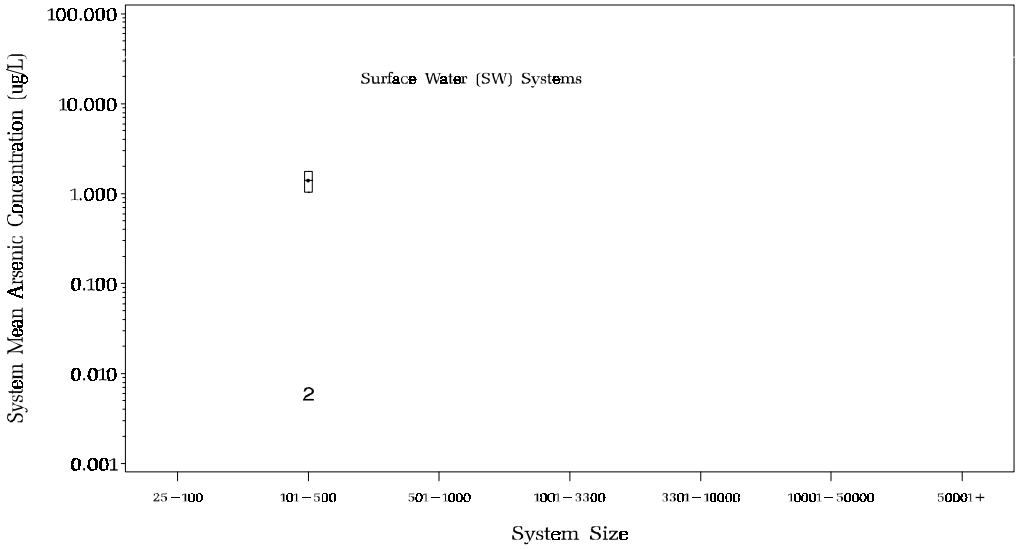


FIGURE 6B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State KS
Number of Systems Indicated Below Boxplot

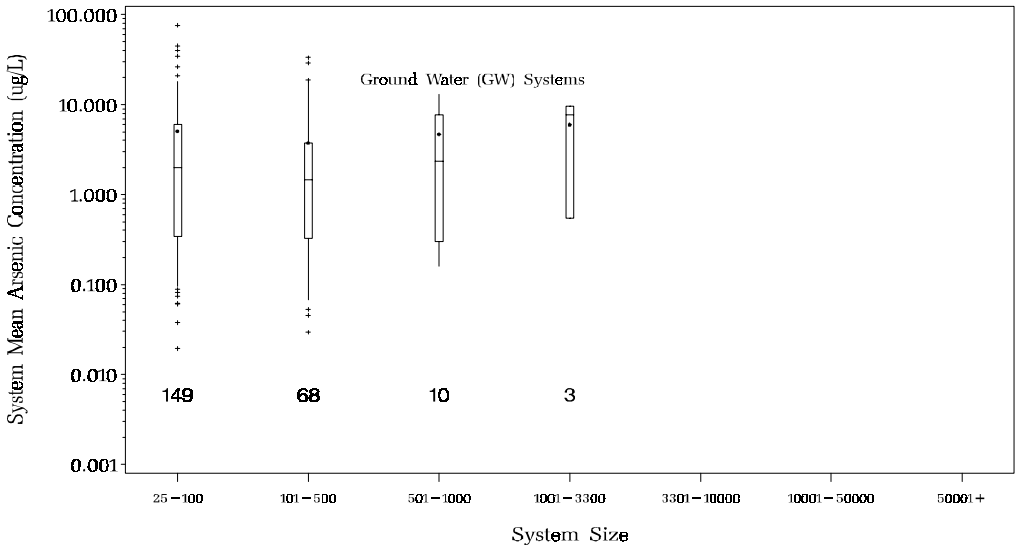


FIGURE 7B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State MI
Number of Systems Indicated Below Boxplot

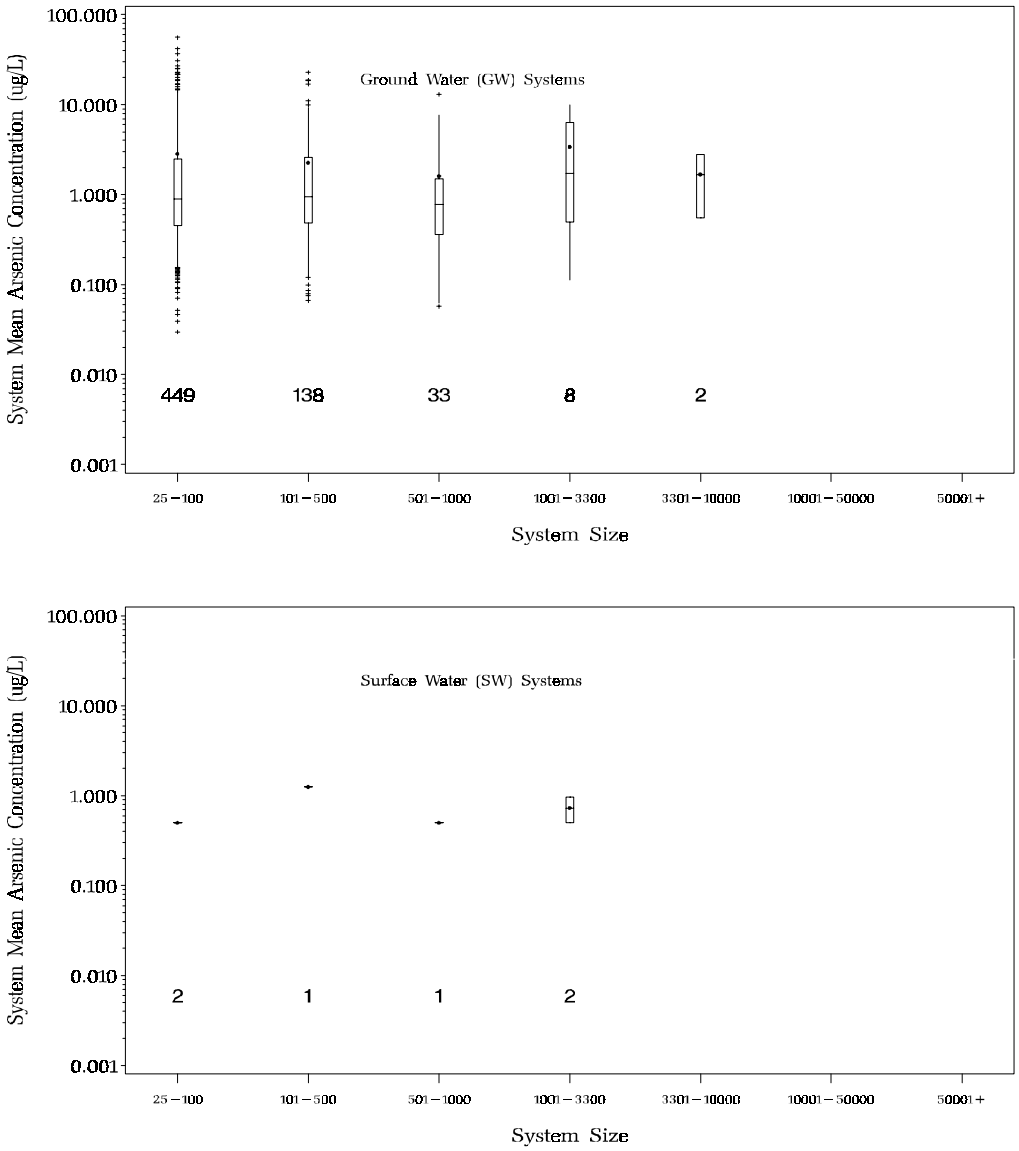


FIGURE 8B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State MN
Number of Systems Indicated Below Boxplot

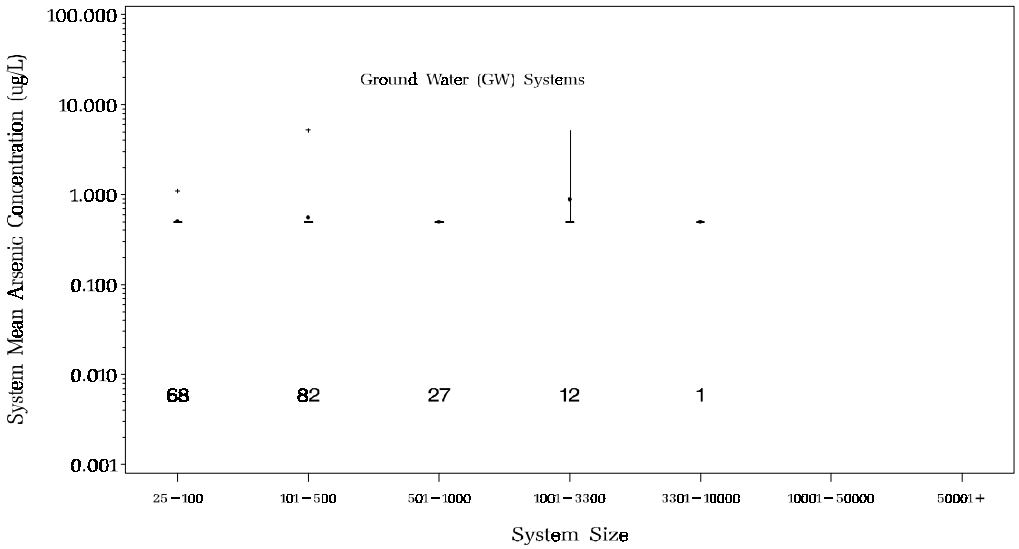


FIGURE 9B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State MO
Number of Systems Indicated Below Boxplot

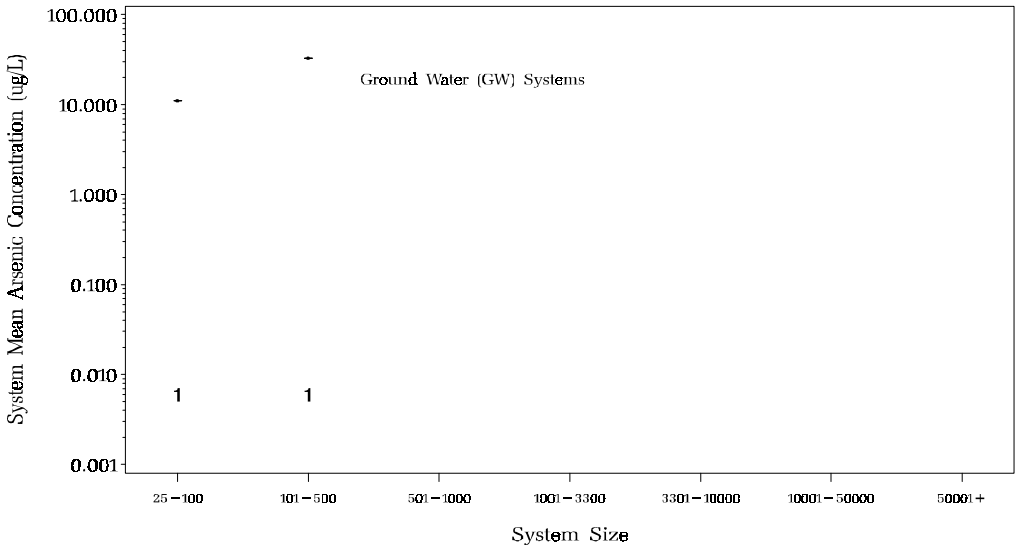
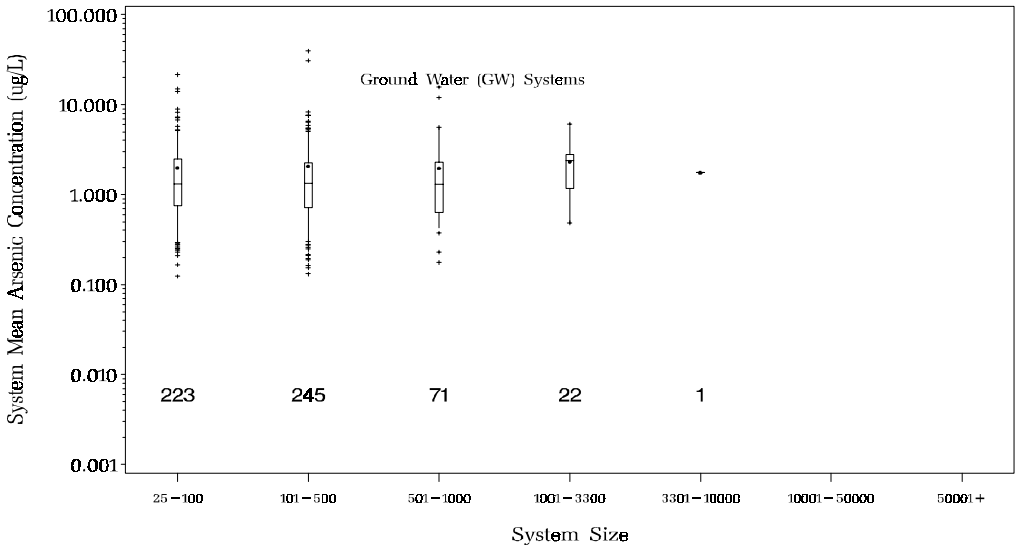


FIGURE 10B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State MT
Number of Systems Indicated Below Boxplot



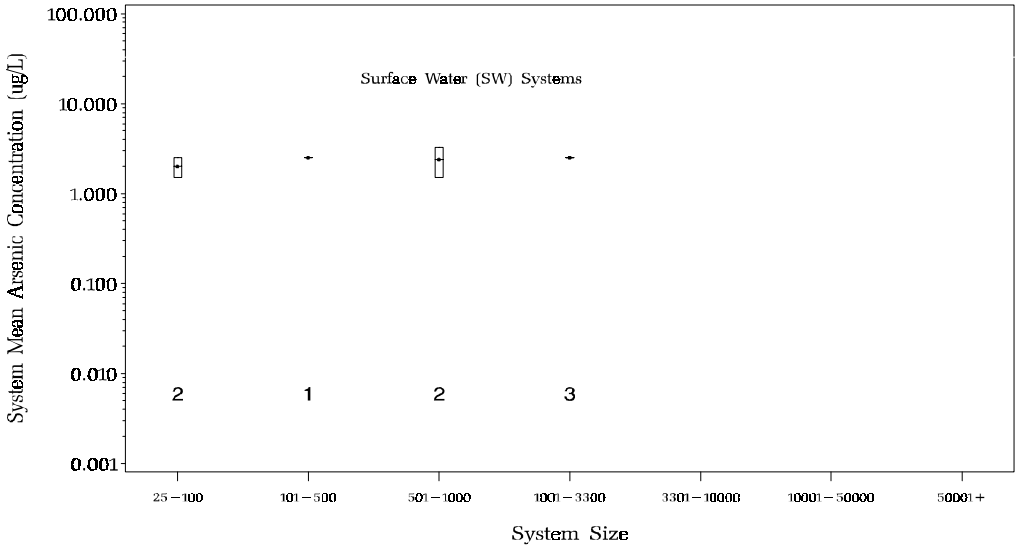
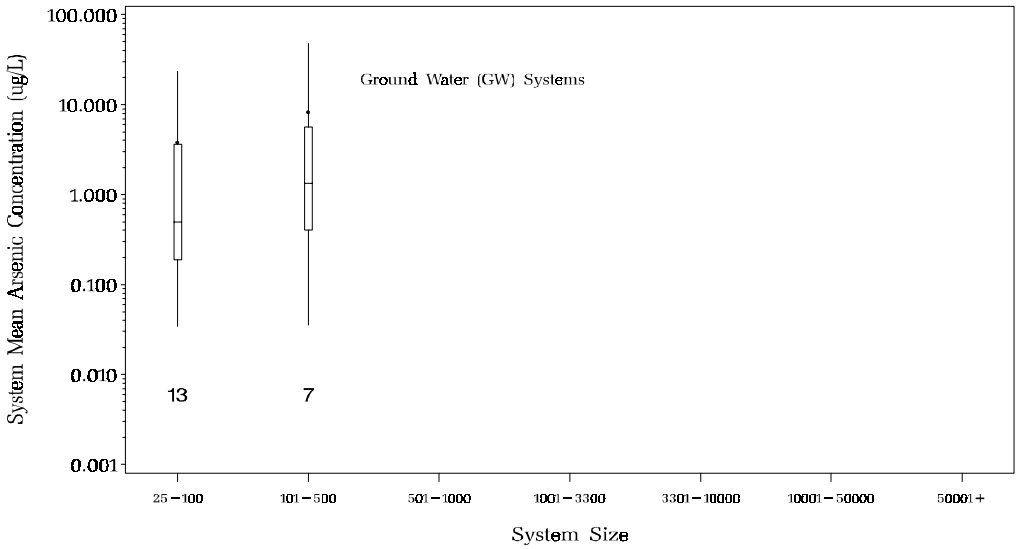


FIGURE 11B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State NC
Number of Systems Indicated Below Boxplot



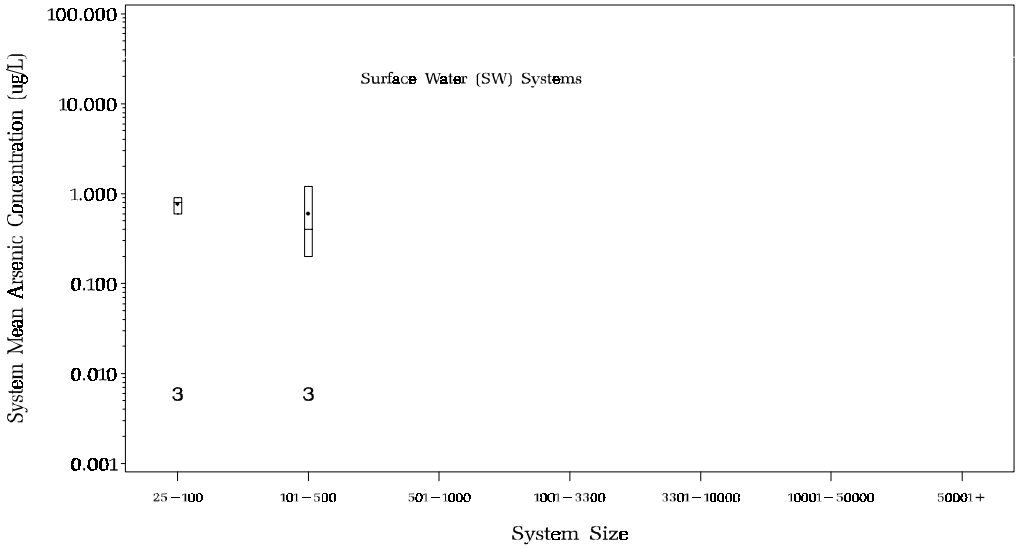
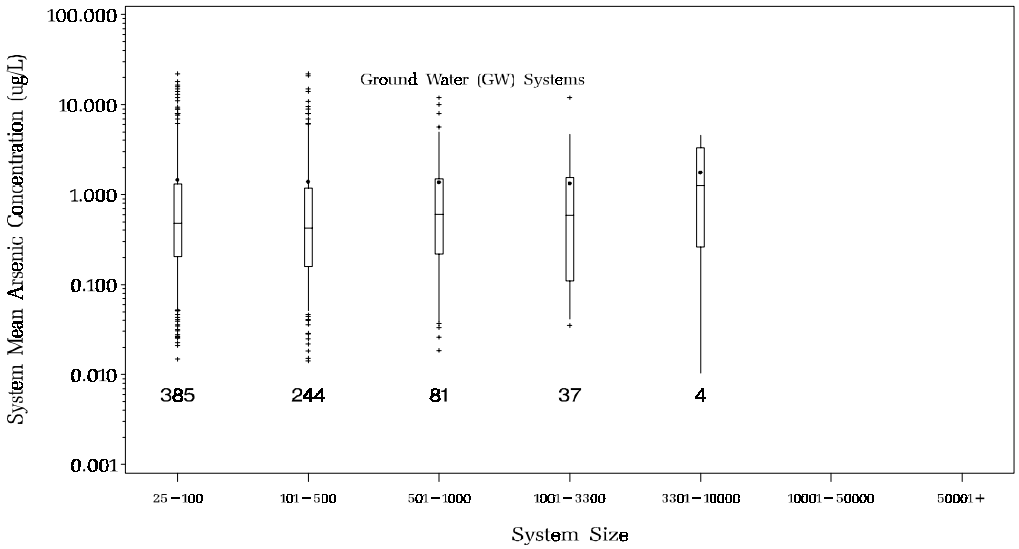


FIGURE 12B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State ND
Number of Systems Indicated Below Boxplot



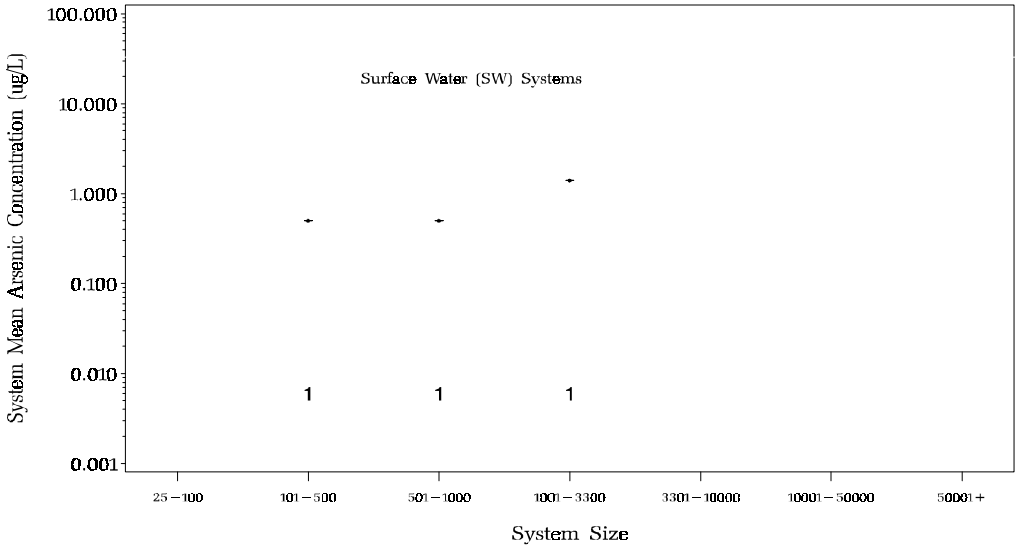
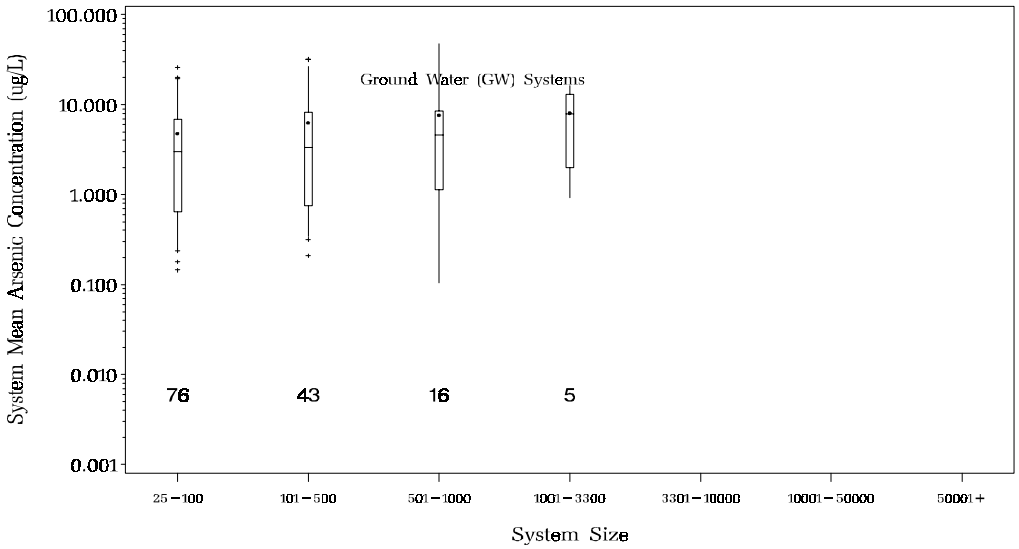


FIGURE 13B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State NJ
Number of Systems Indicated Below Boxplot



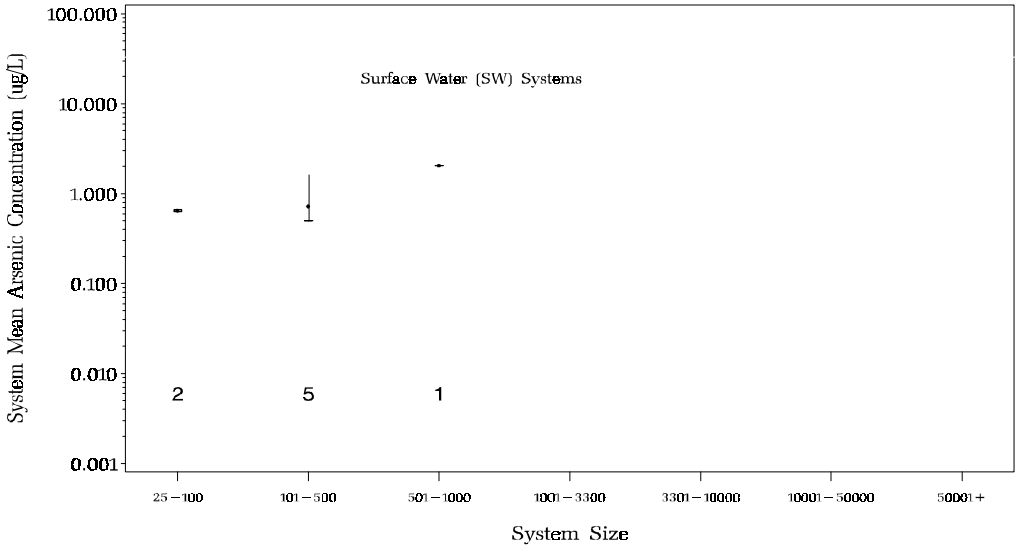


FIGURE 14B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State NM
Number of Systems Indicated Below Boxplot

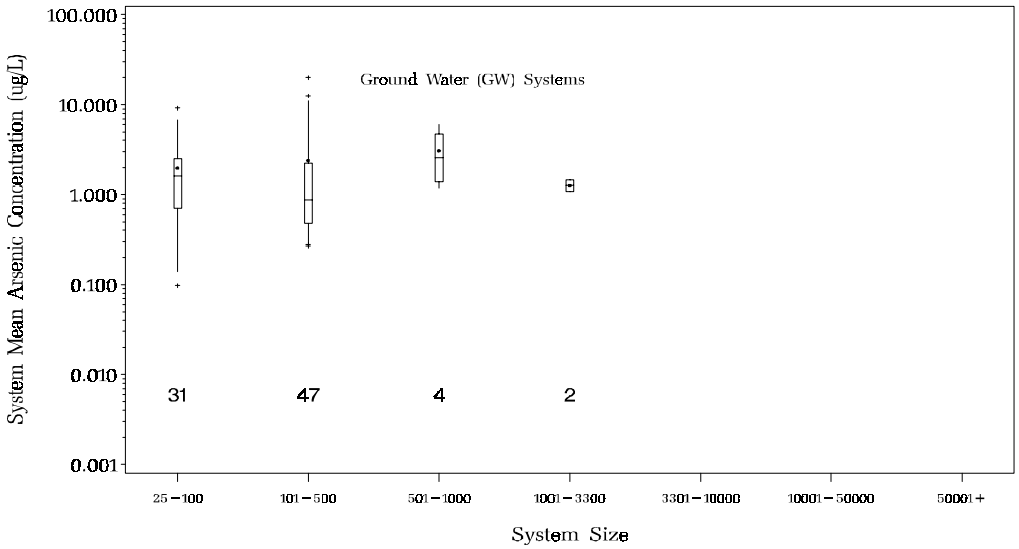
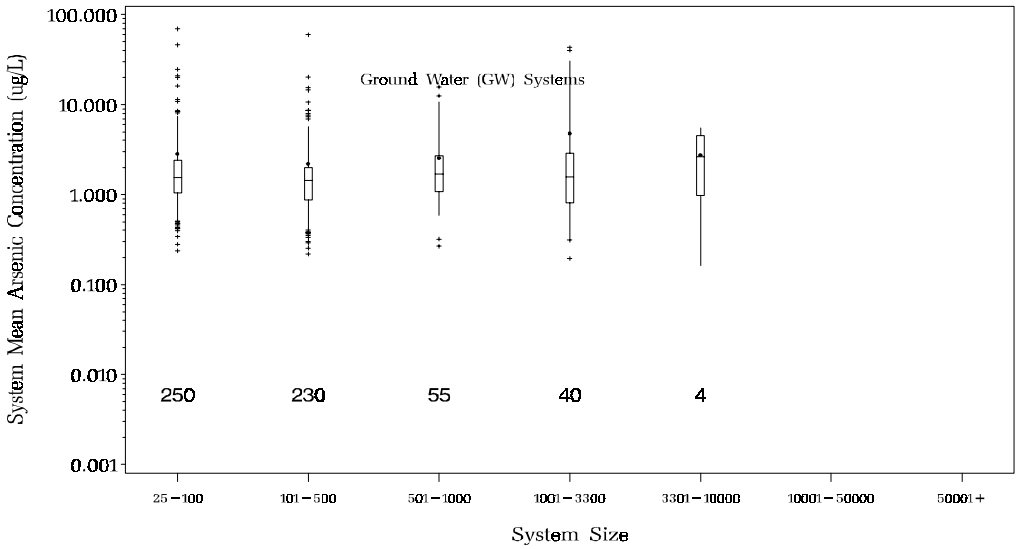


FIGURE 15B. Boxplots of System Means by System Size for Non – Transient Non – Community Water Systems in State OR Number of Systems Indicated Below Boxplot



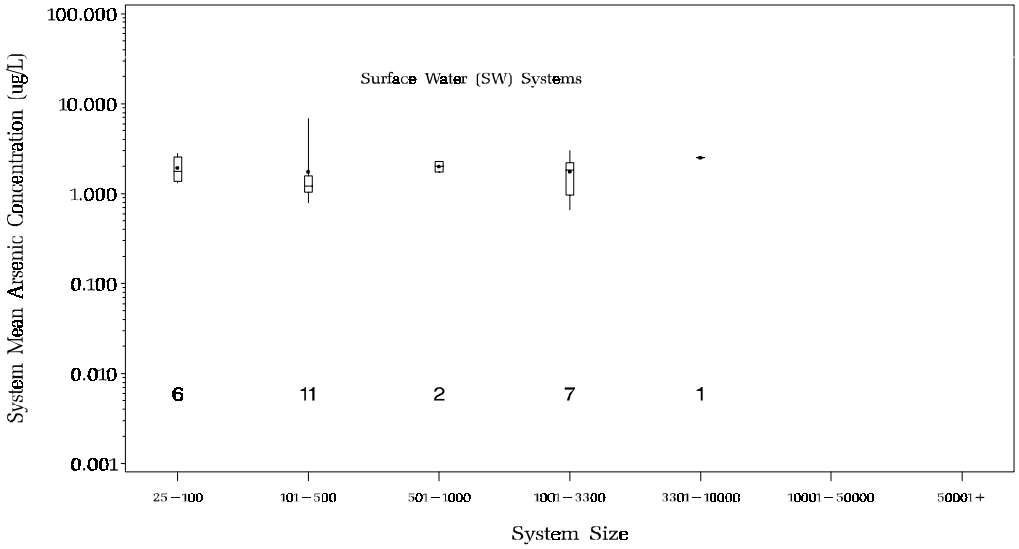


FIGURE 16B. Boxplots of System Means by System Size for
Non – Transient Non – Community Water Systems in State TX
Number of Systems Indicated Below Boxplot

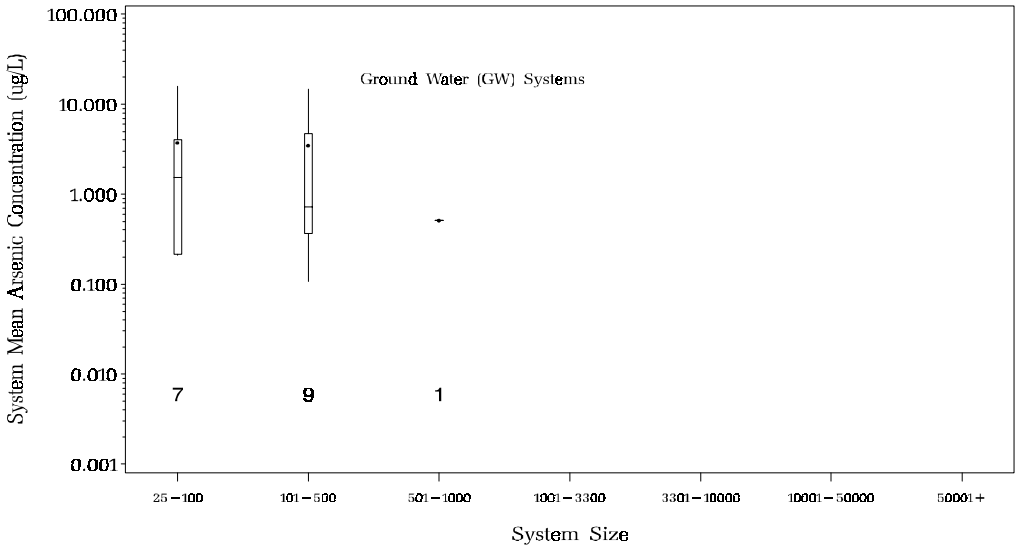


FIGURE 17B. Boxplots of System Means by System Size for Non – Transient Non – Community Water Systems in State UT Number of Systems Indicated Below Boxplot

Appendix B-3
Lognormal Probability Plots of System Means
(Only includes States with 5 or more systems that are not completely censored)

This page intentionally left blank

Figure B−1: System means of CWS GW arsenic concentrations for AK, Log−normal probability plot

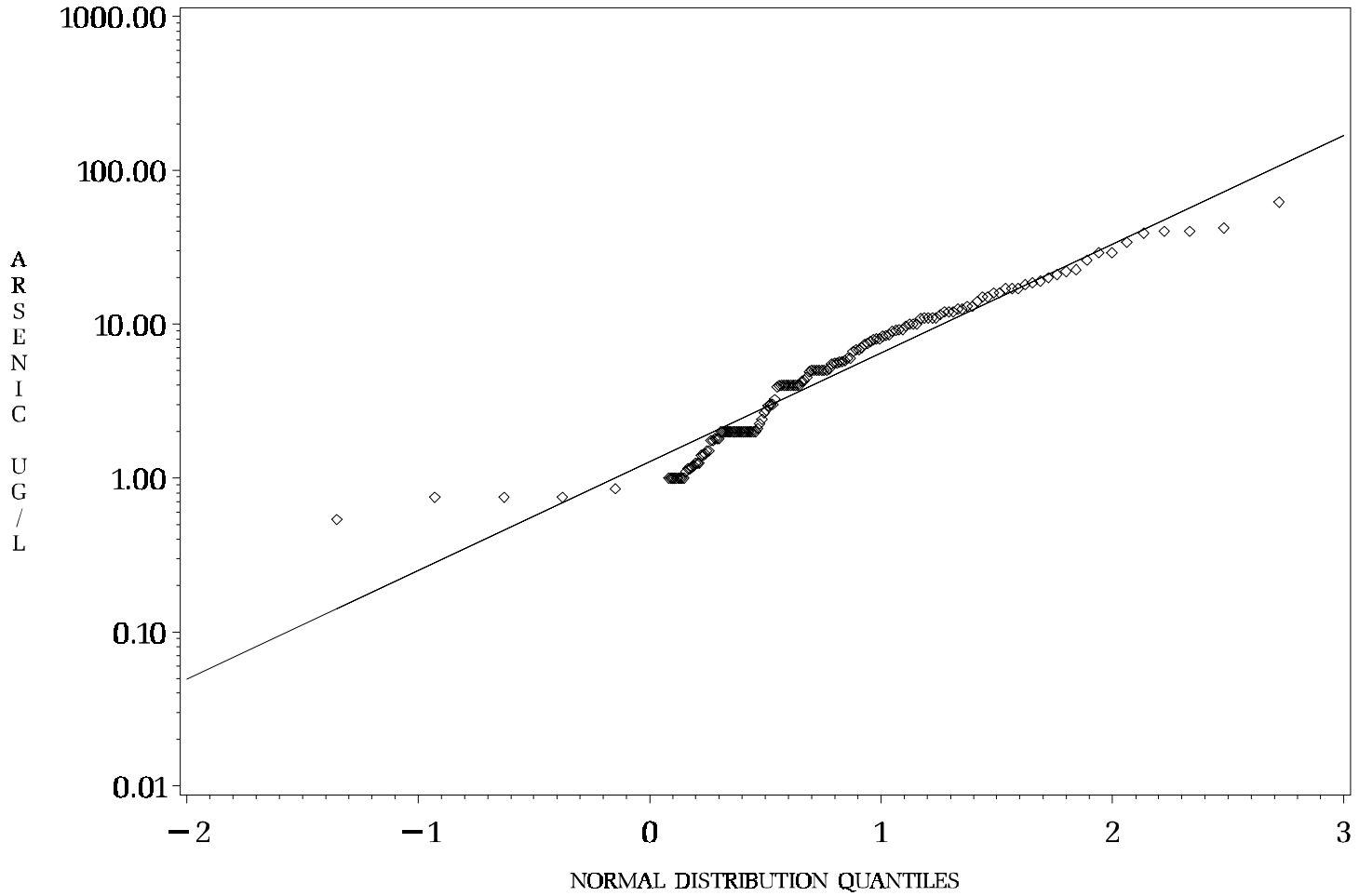


Figure B−2: System means of CWS GW arsenic concentrations for AL, Log−normal probability plot

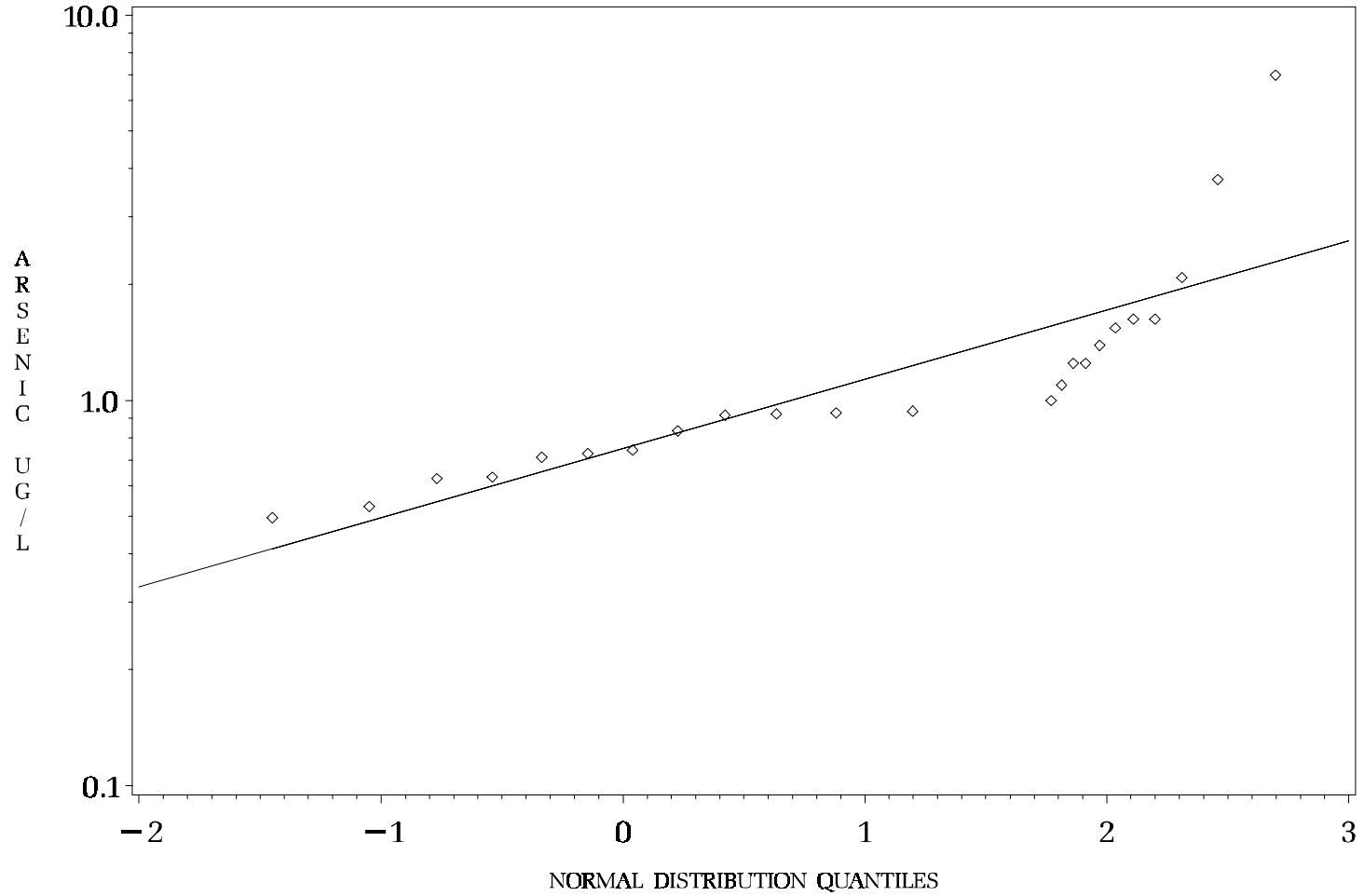


Figure B−3: System means of CWS GW arsenic concentrations for AZ, Log−normal probability plot

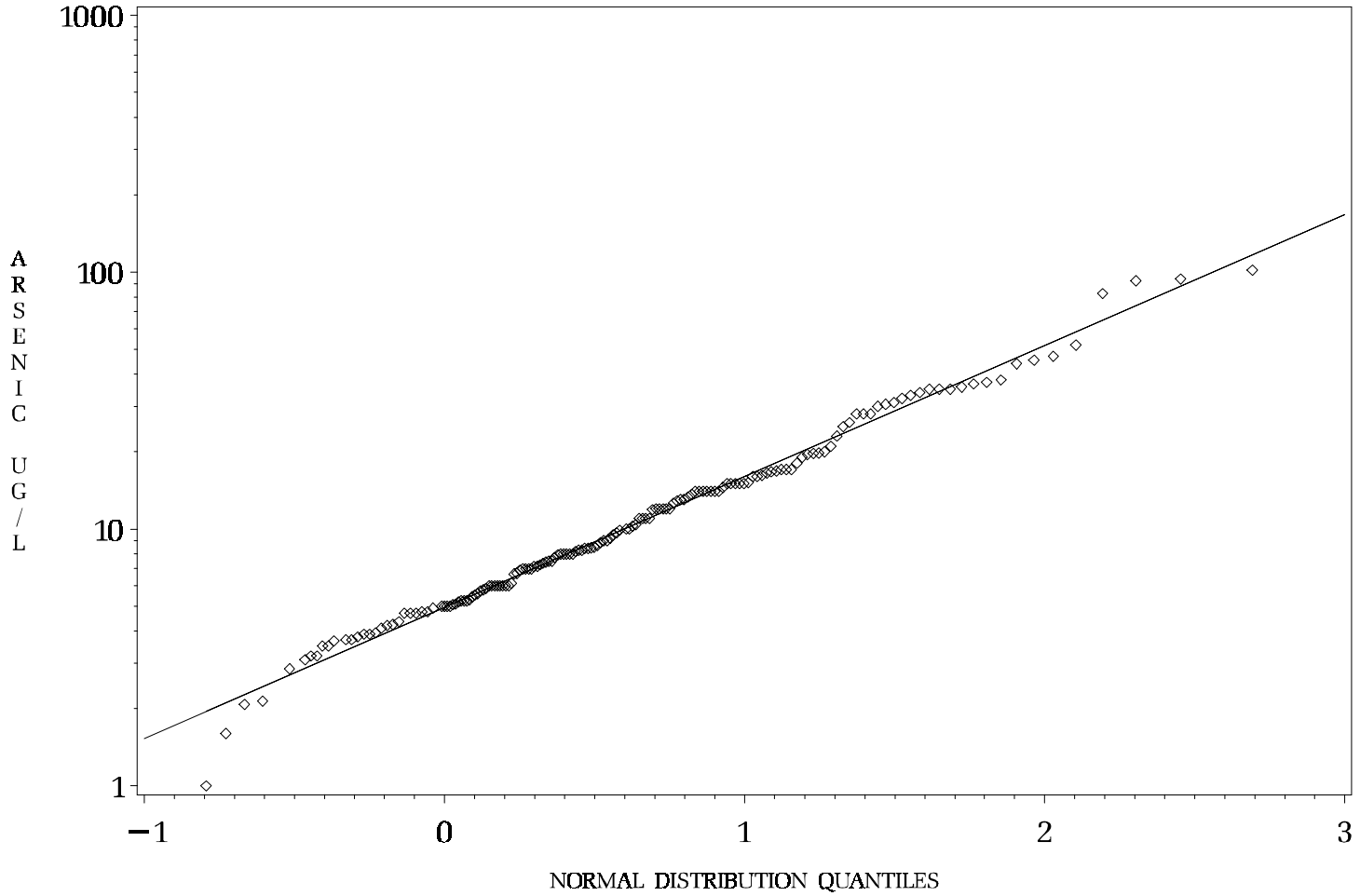


Figure B−4: System means of CWS GW arsenic concentrations for CA, Log−normal probability plot

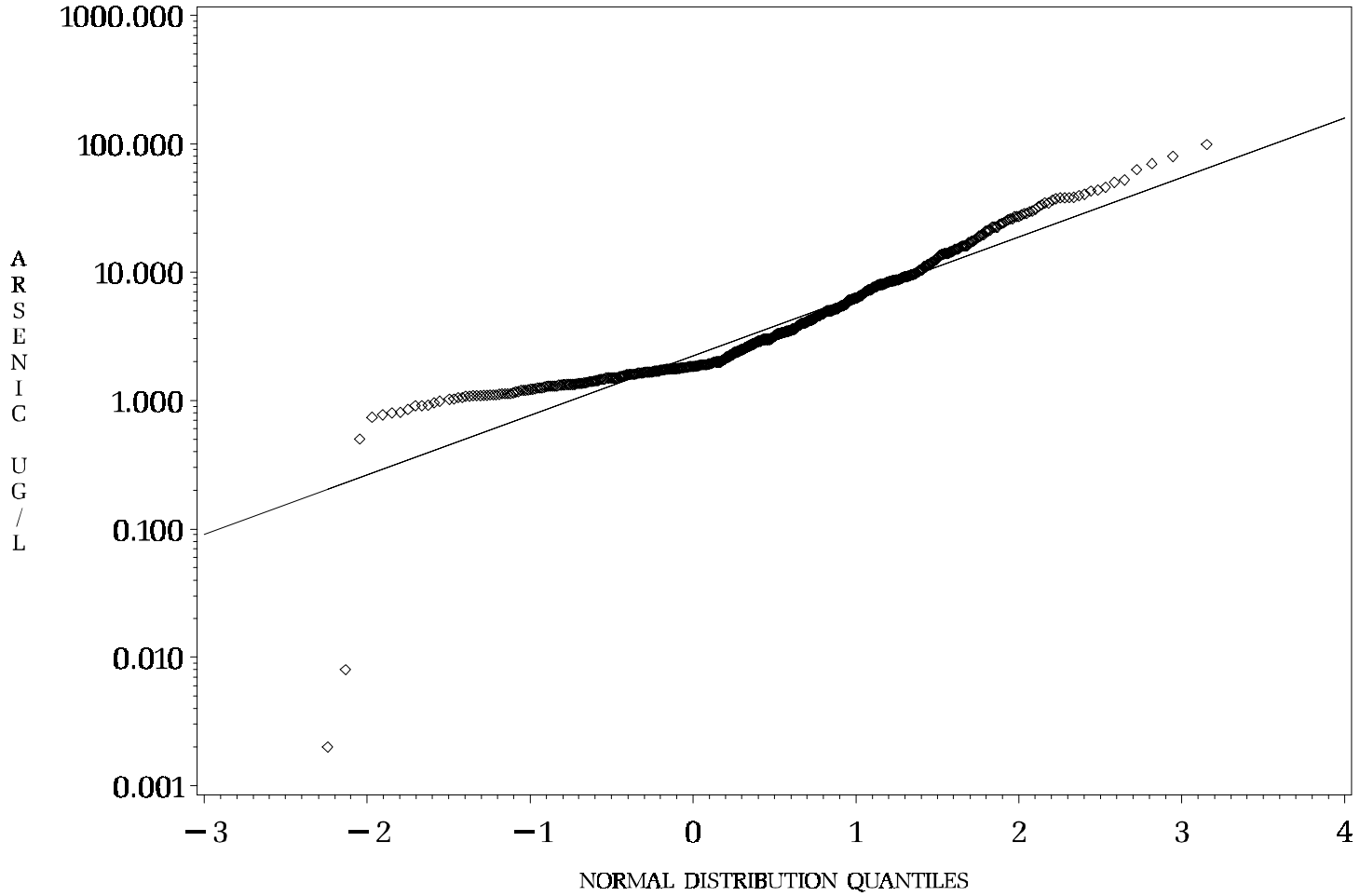


Figure B−5: System means of CWS GW arsenic concentrations for IL, Log−normal probability plot

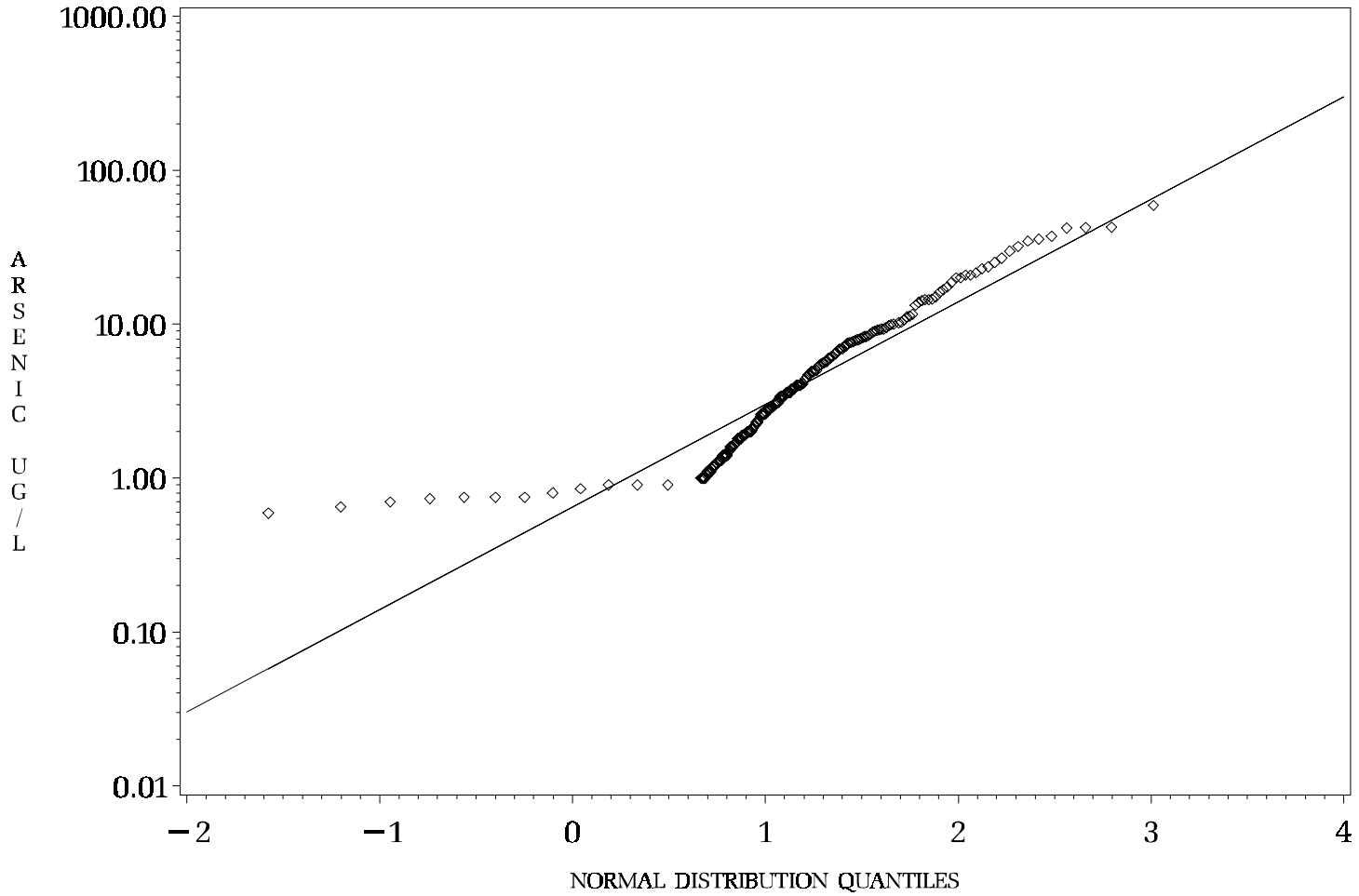


Figure B−6: System means of CWS GW arsenic concentrations for IN, Log−normal probability plot

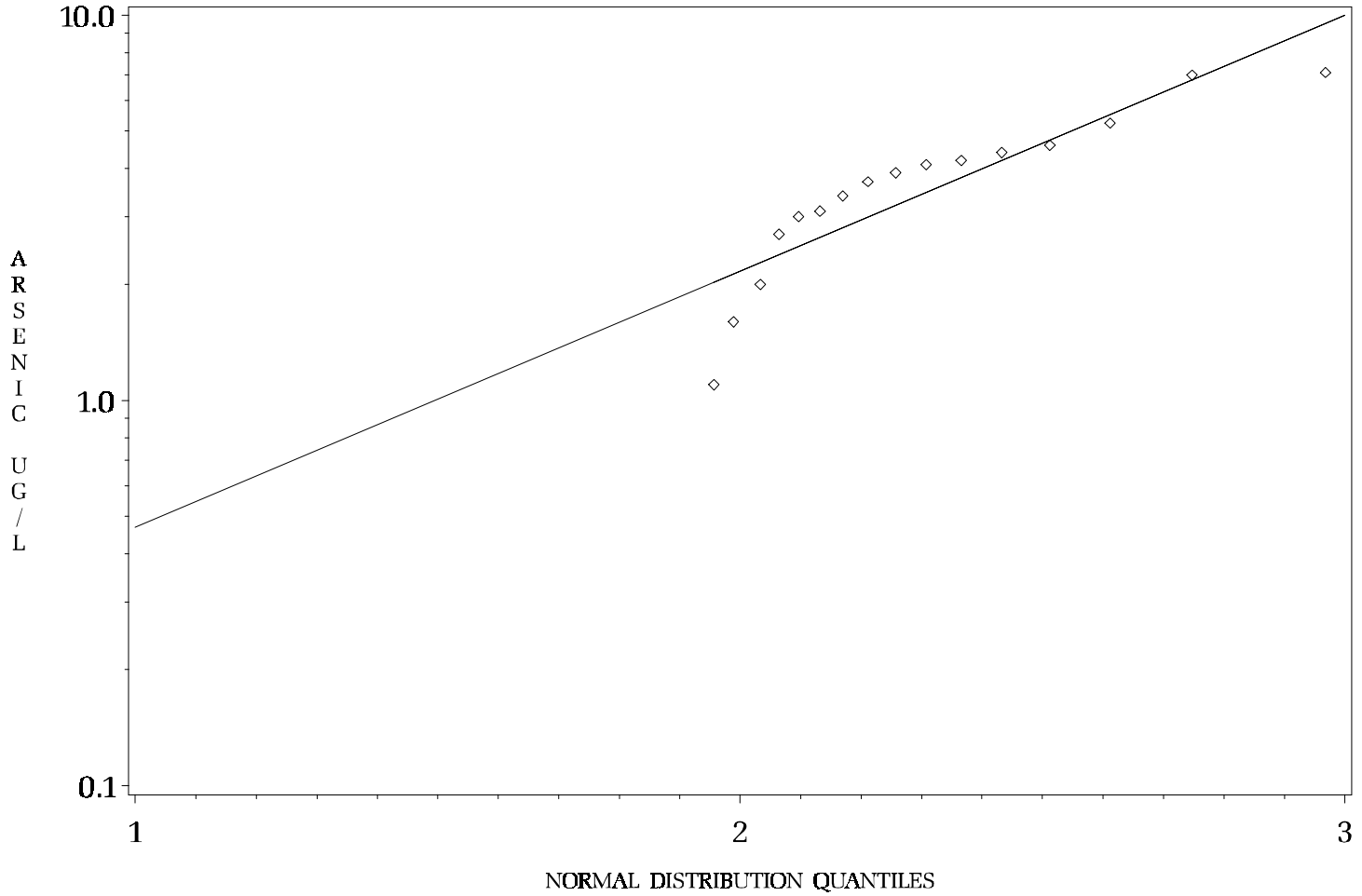


Figure B−7: System means of CWS GW arsenic concentrations for KS, Log−normal probability plot

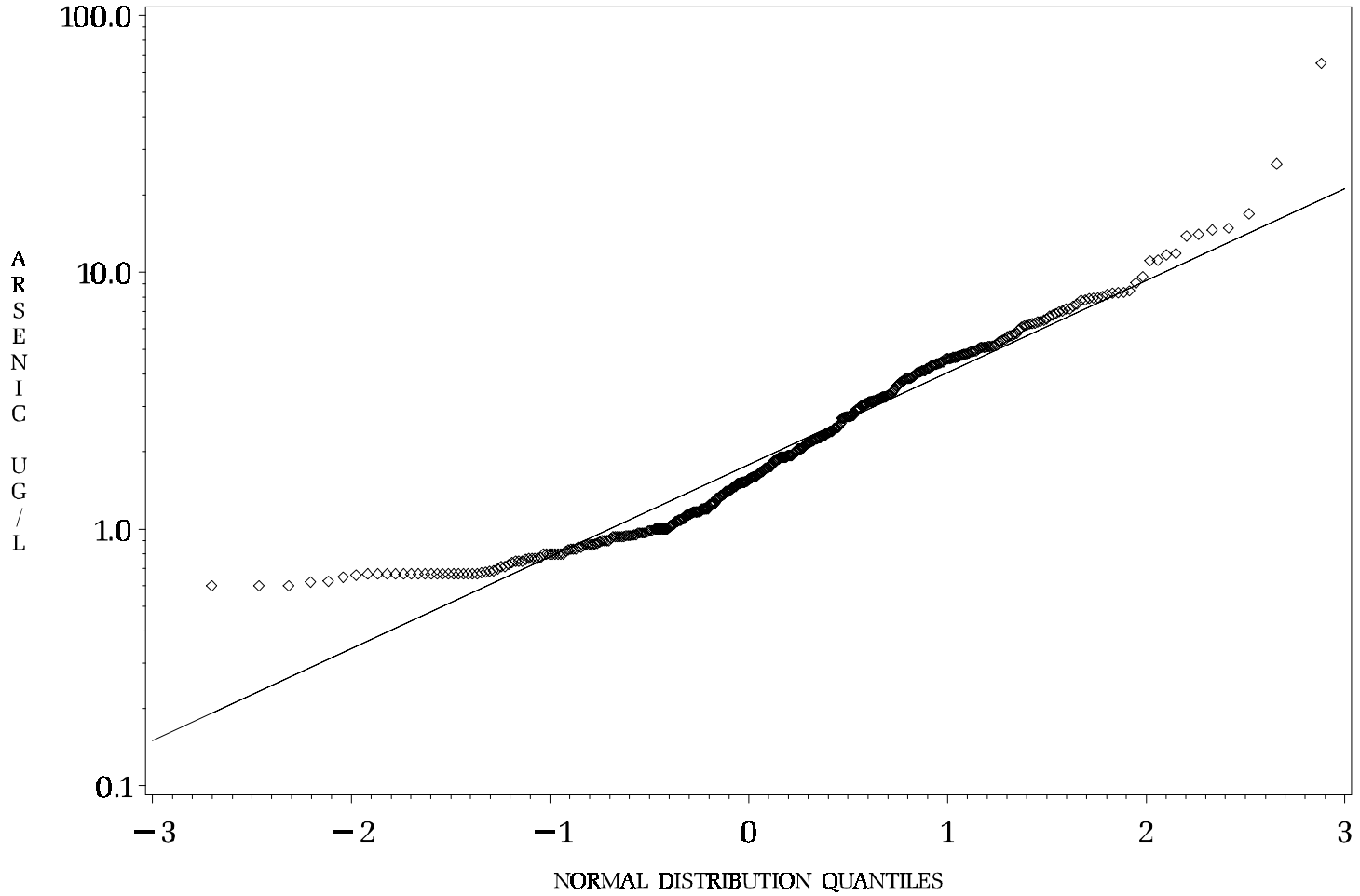


Figure B−8: System means of CWS GW arsenic concentrations for KY, Log−normal probability plot

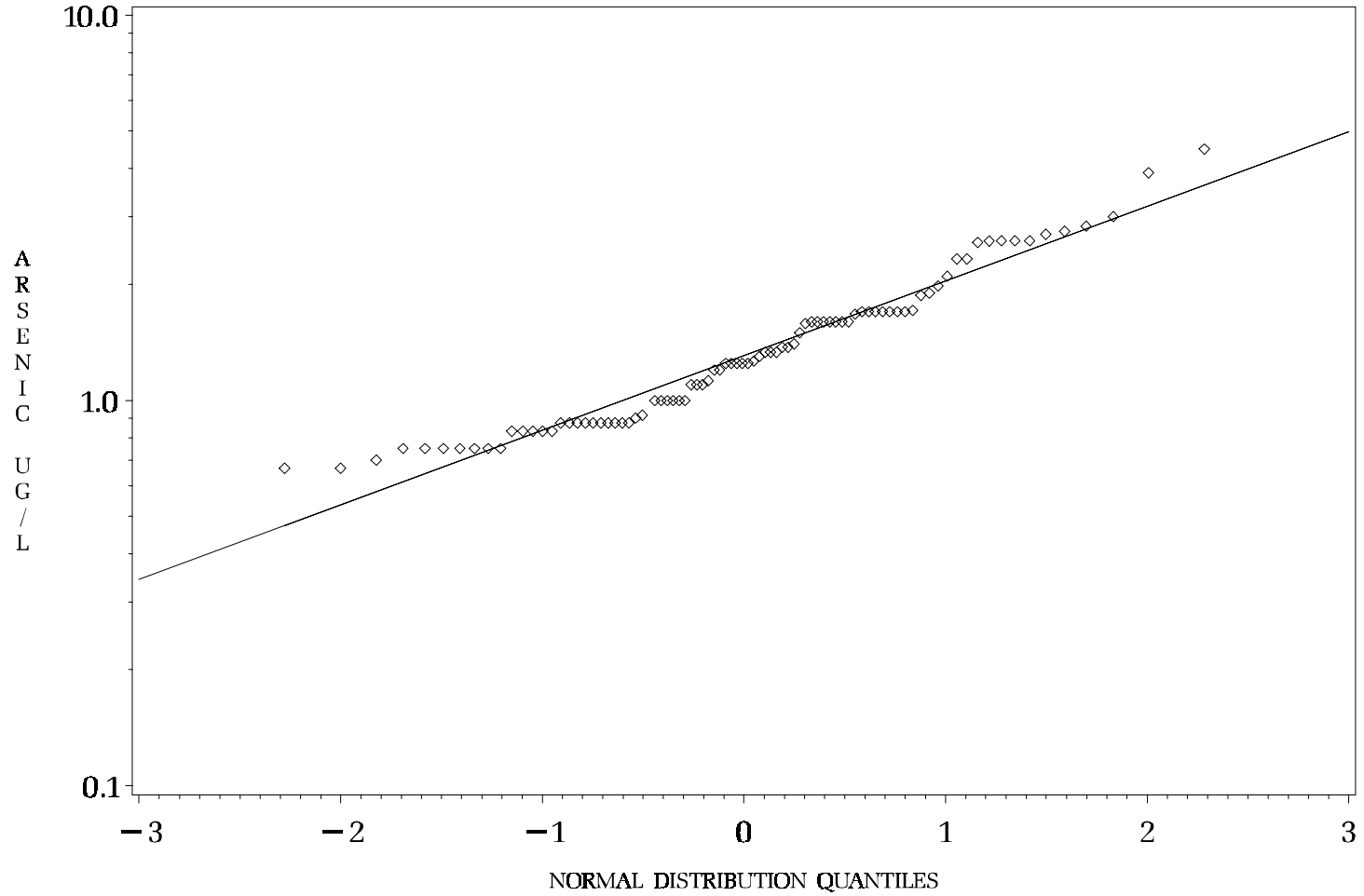


Figure B−9: System means of CWS GW arsenic concentrations for ME, Log−normal probability plot

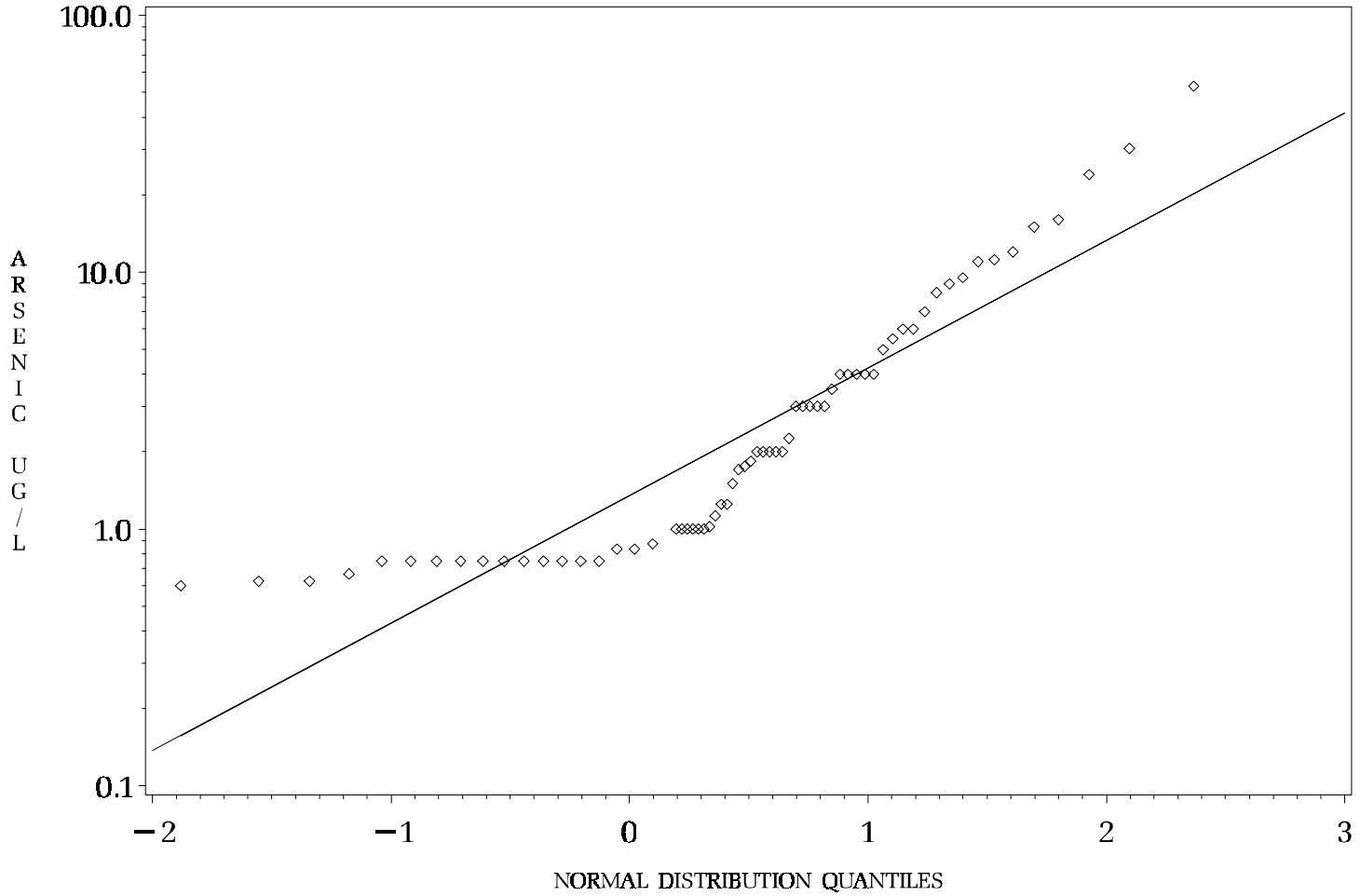


Figure B−10: System means of CWS GW arsenic concentrations for MI, Log−normal probability plot

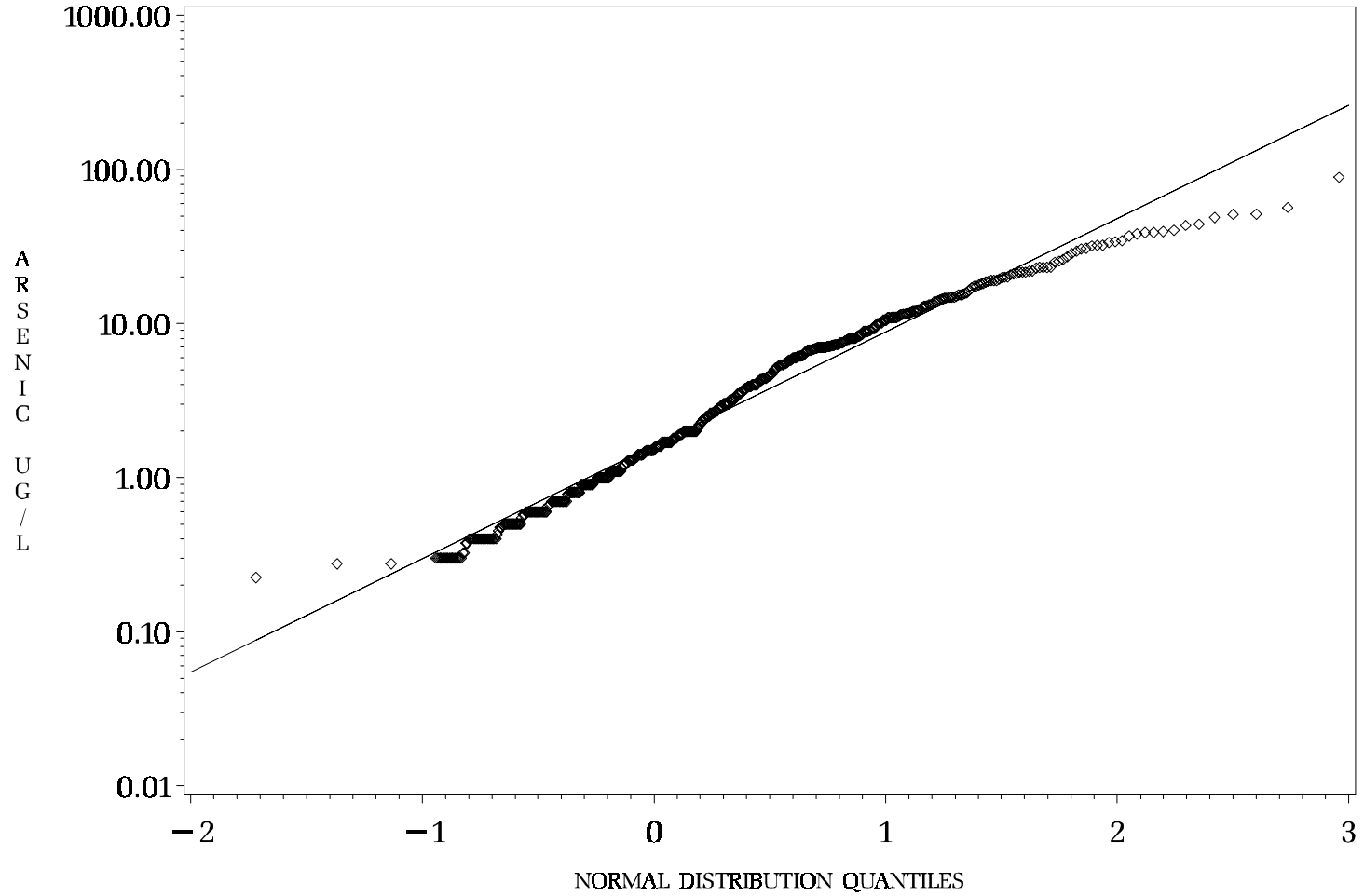


Figure B−11: System means of CWS GW arsenic concentrations for MN, Log−normal probability plot

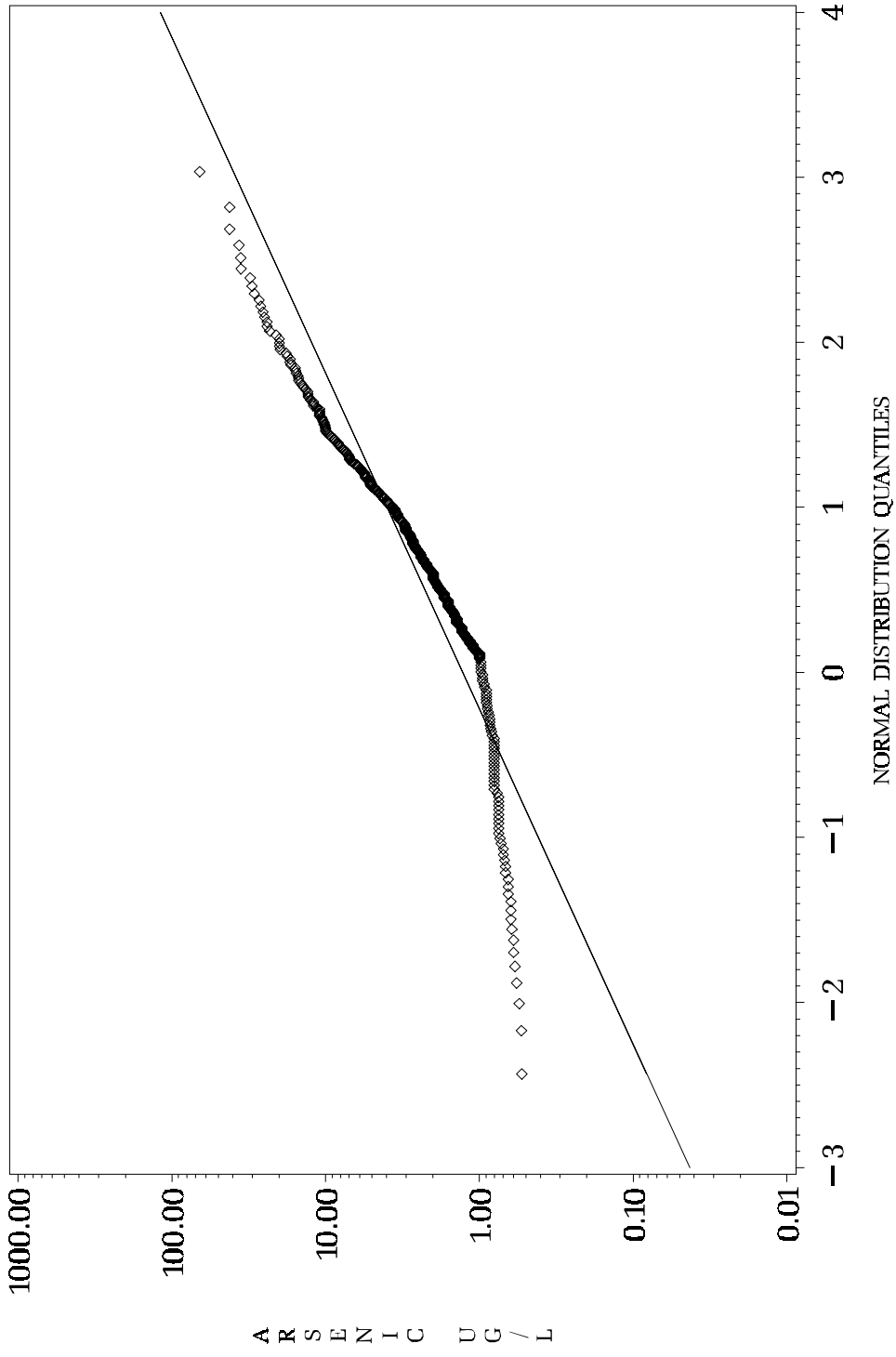


Figure B−12: System means of CWS GW arsenic concentrations for MO, Log−normal probability plot

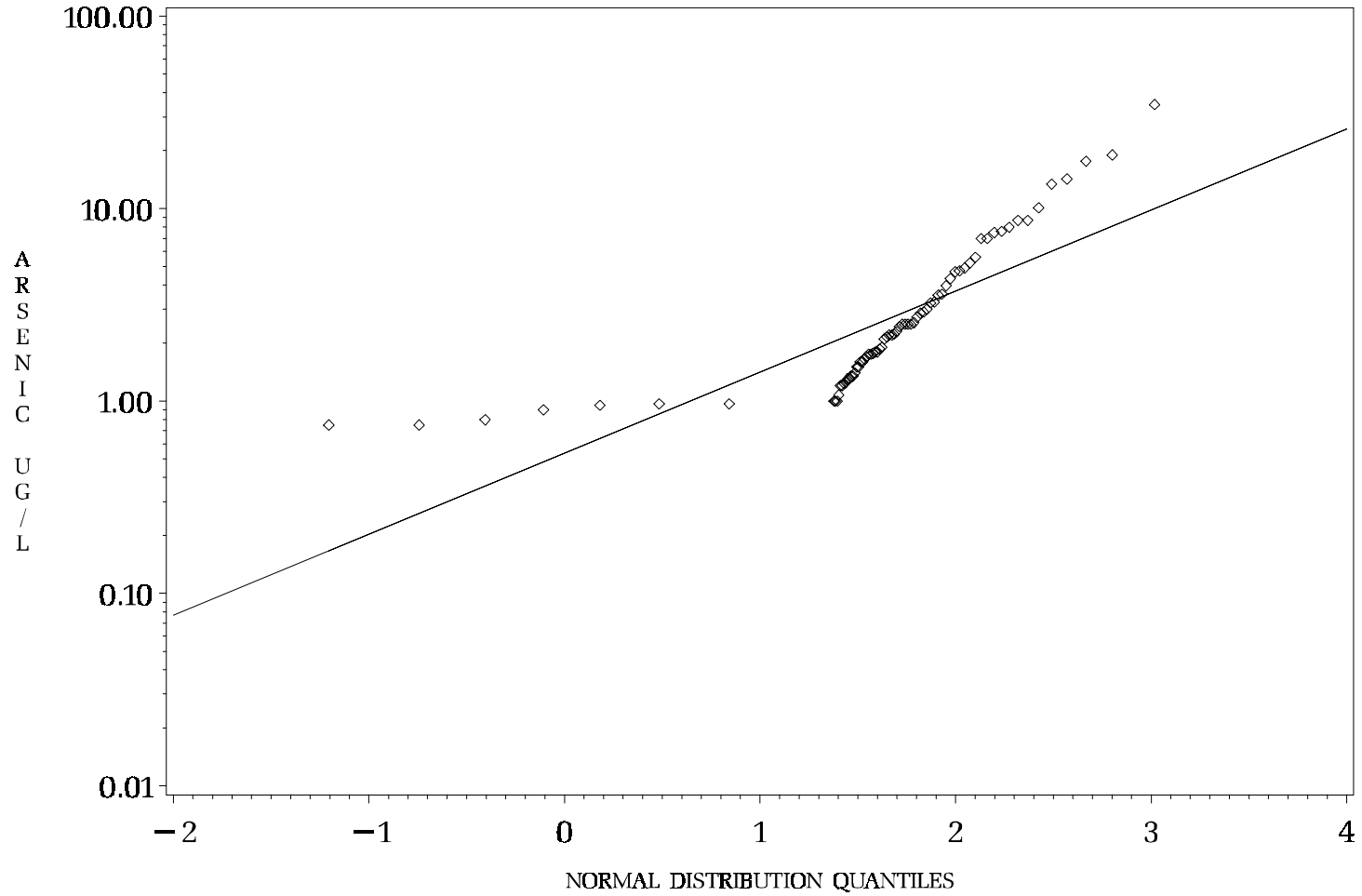


Figure B−13: System means of CWS GW arsenic concentrations for MT, Log−normal probability plot

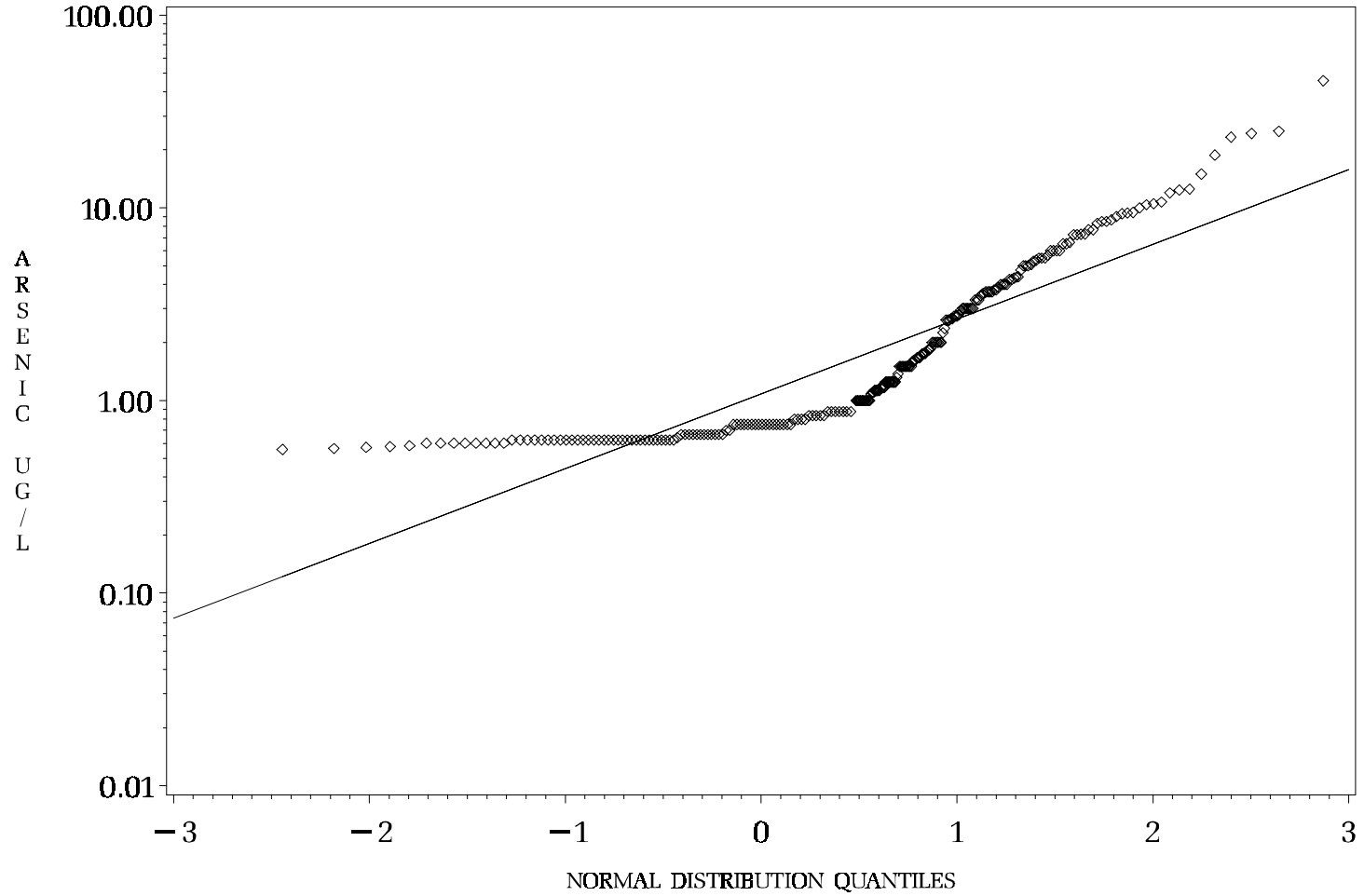


Figure B−14: System means of CWS GW arsenic concentrations for NC, Log−normal probability plot

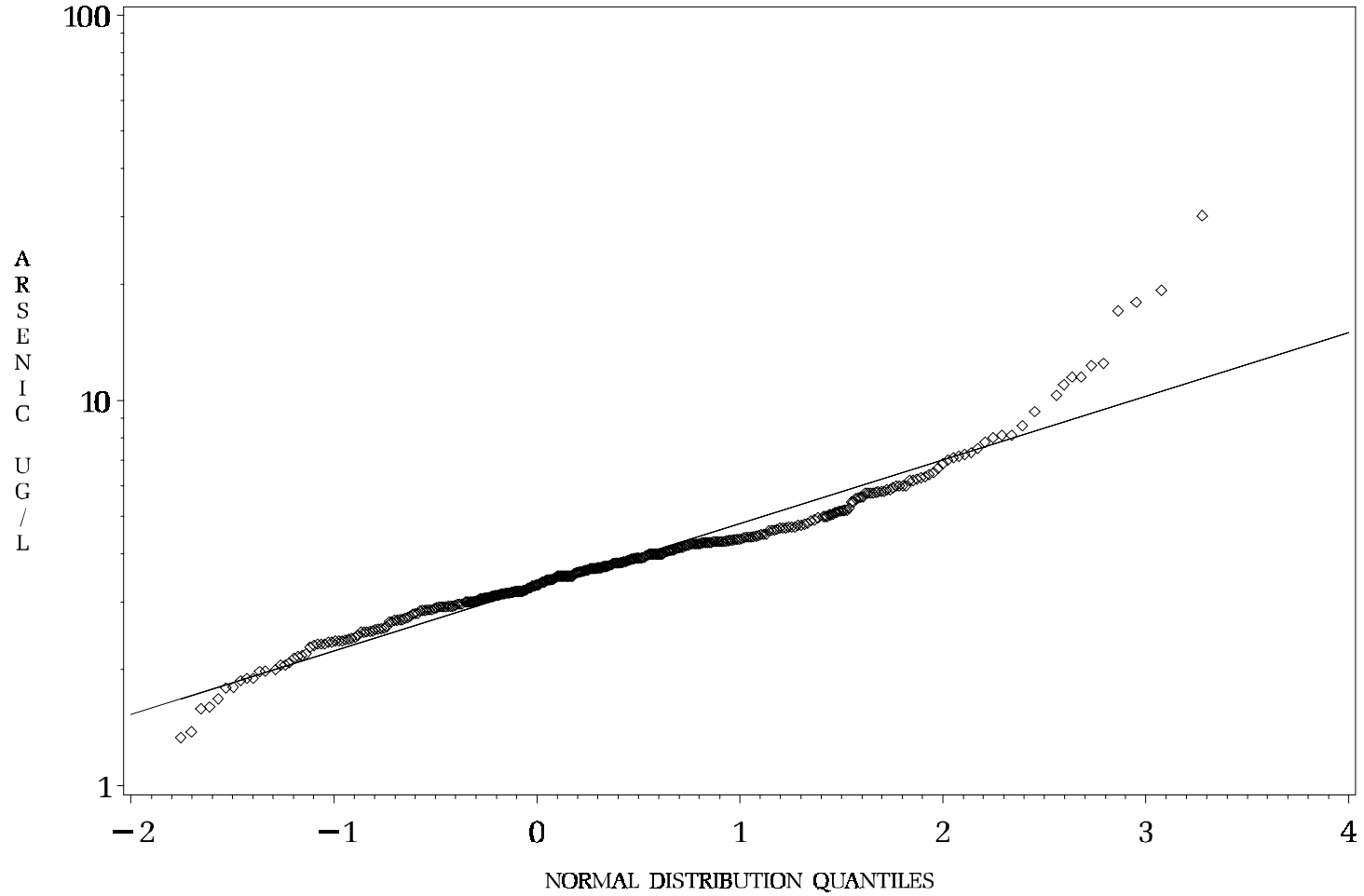


Figure B−15: System means of CWS GW arsenic concentrations for ND, Log−normal probability plot

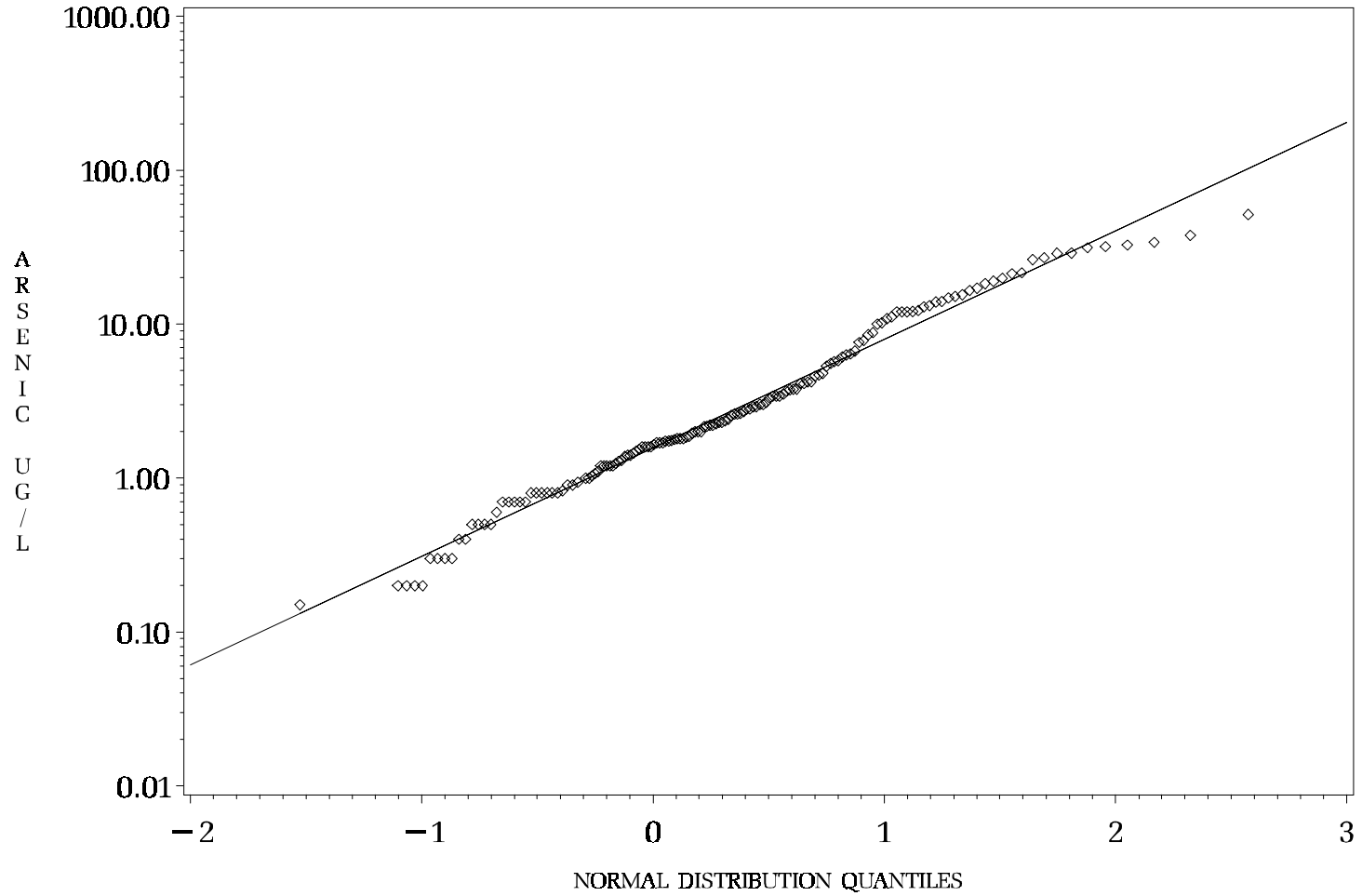


Figure B−16: System means of CWS GW arsenic concentrations for NH, Log−normal probability plot

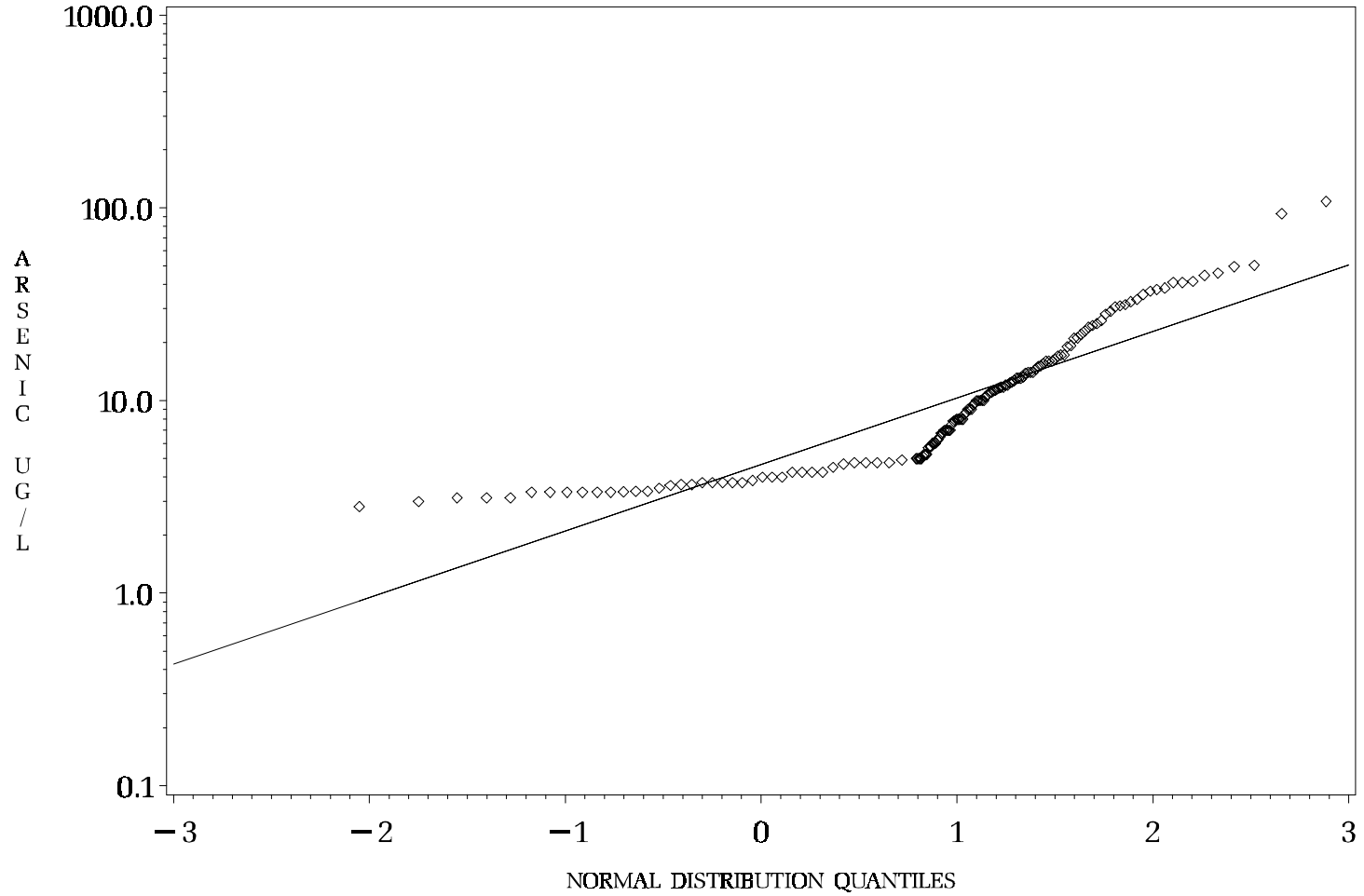


Figure B−17: System means of CWS GW arsenic concentrations for NJ, Log−normal probability plot

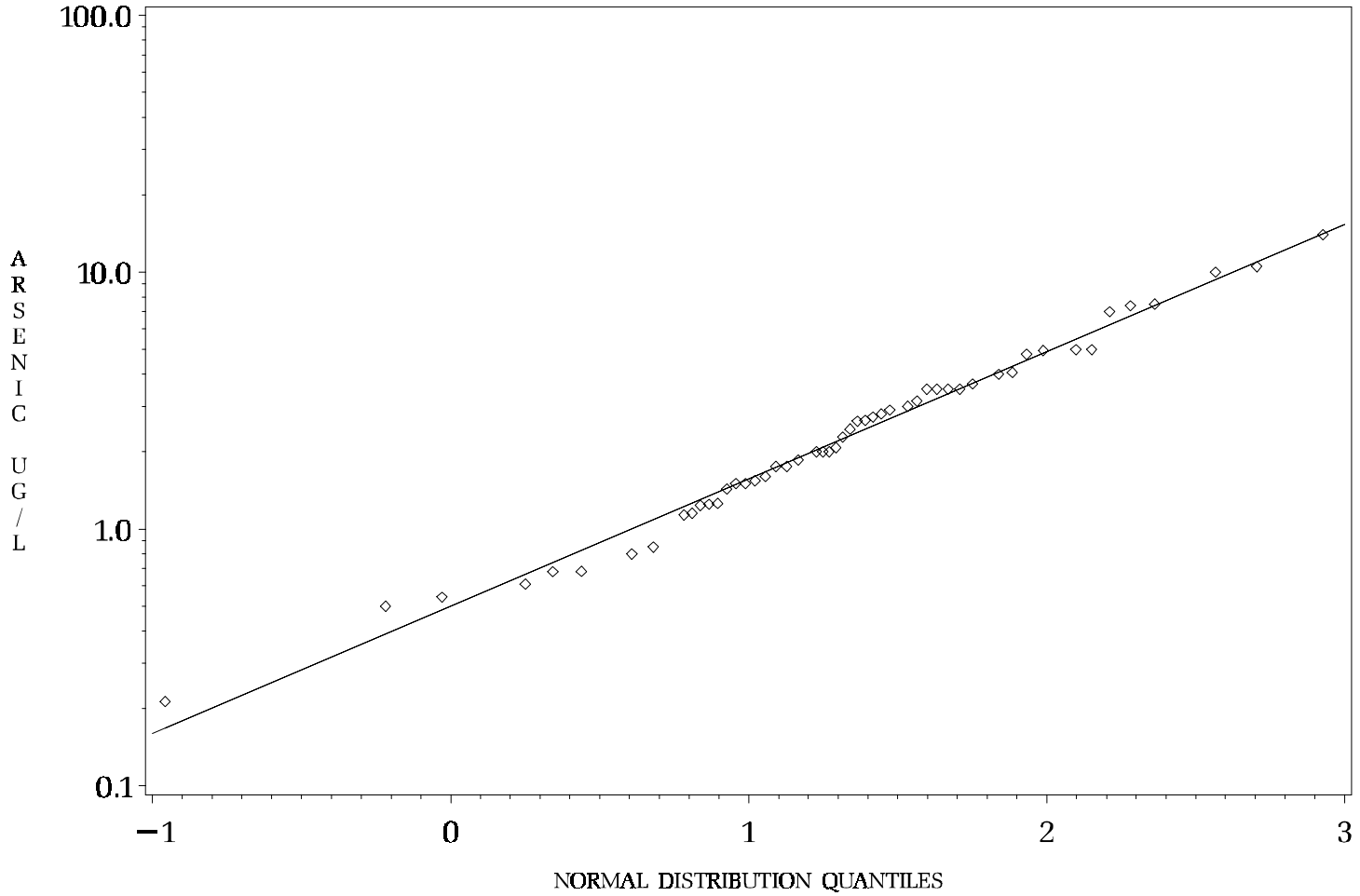


Figure B−18: System means of CWS GW arsenic concentrations for NM, Log−normal probability plot

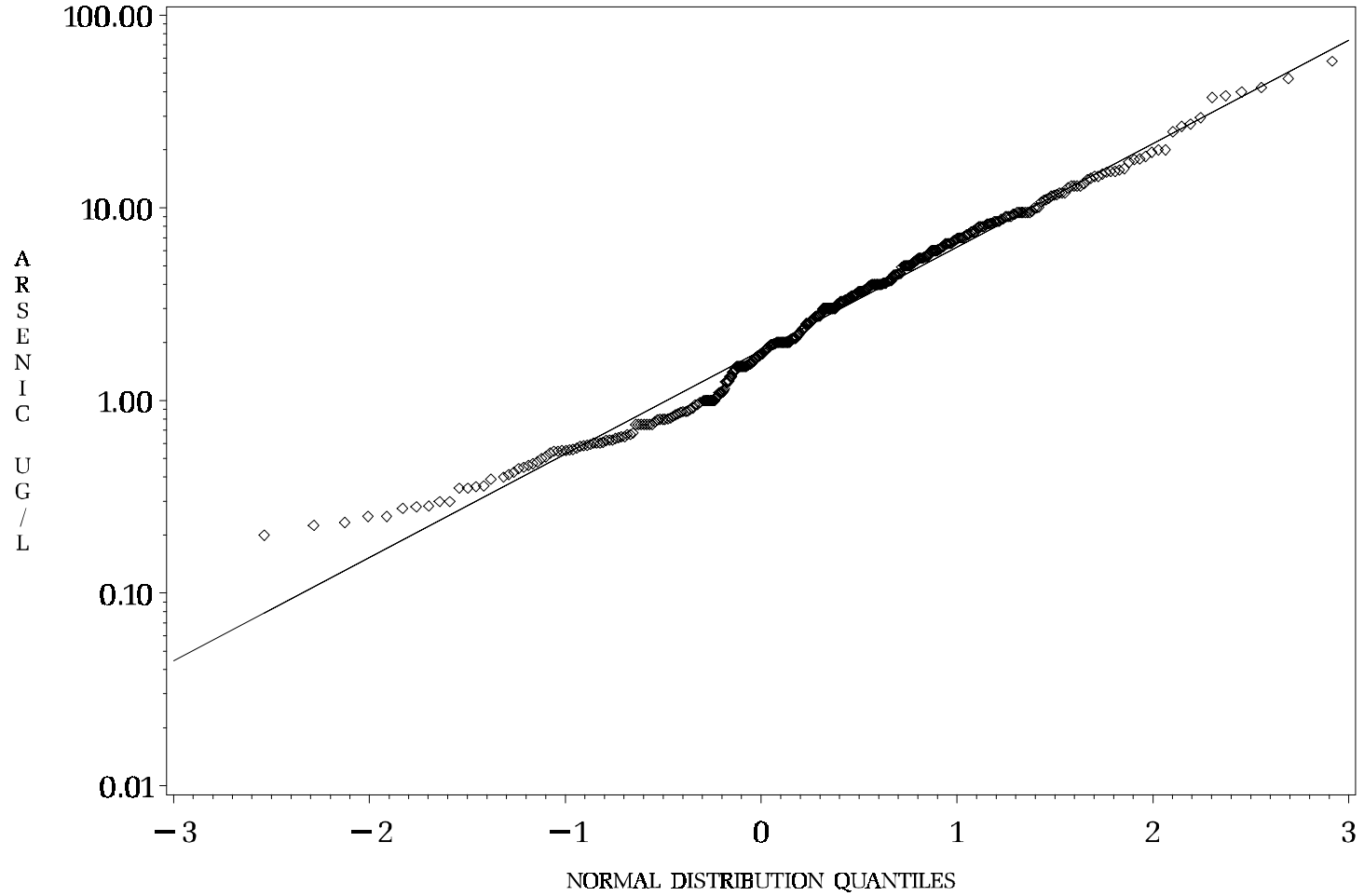


Figure B−19: System means of CWS GW arsenic concentrations for NV, Log−normal probability plot

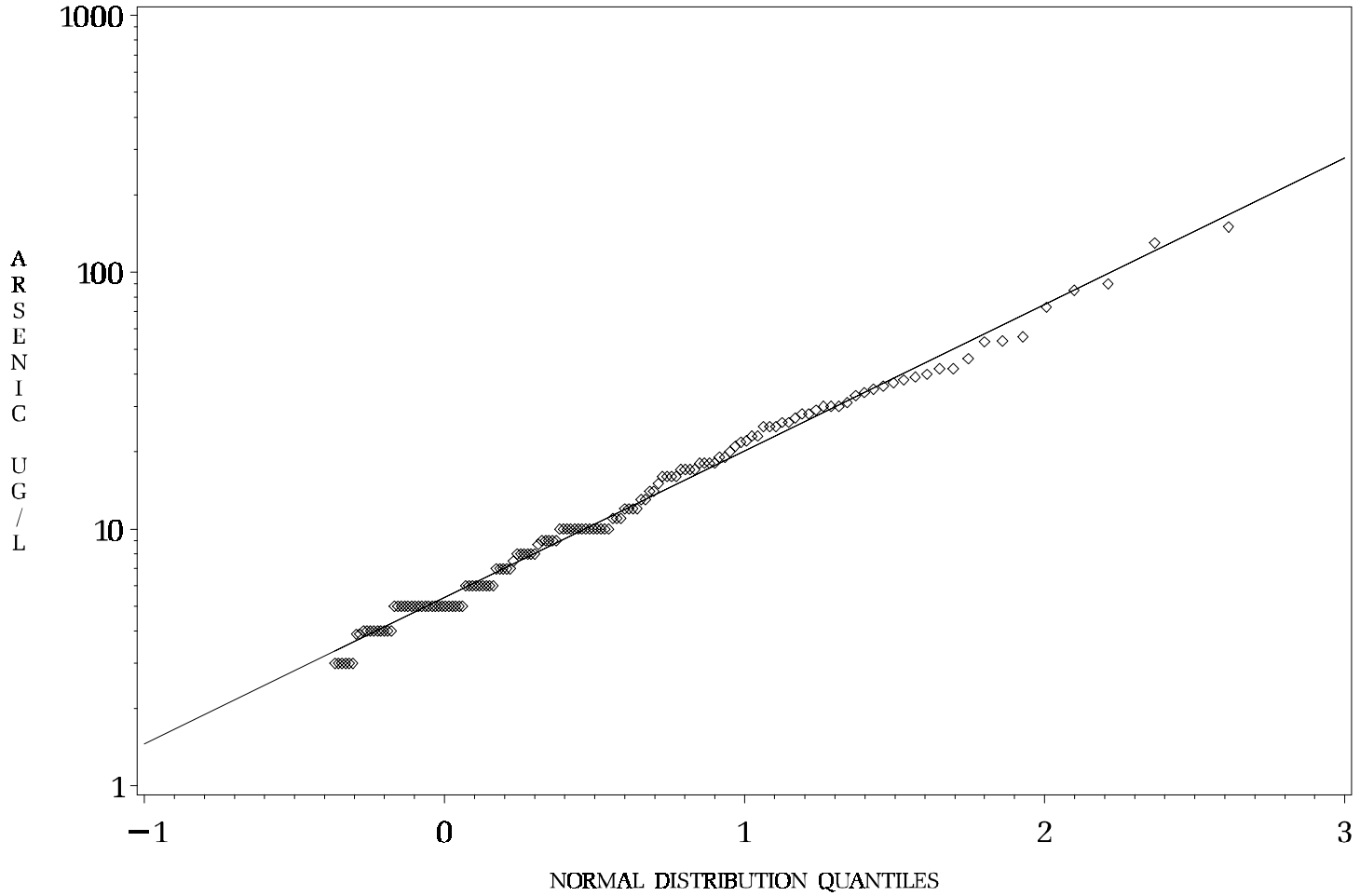


Figure B−20: System means of CWS GW arsenic concentrations for OH, Log−normal probability plot

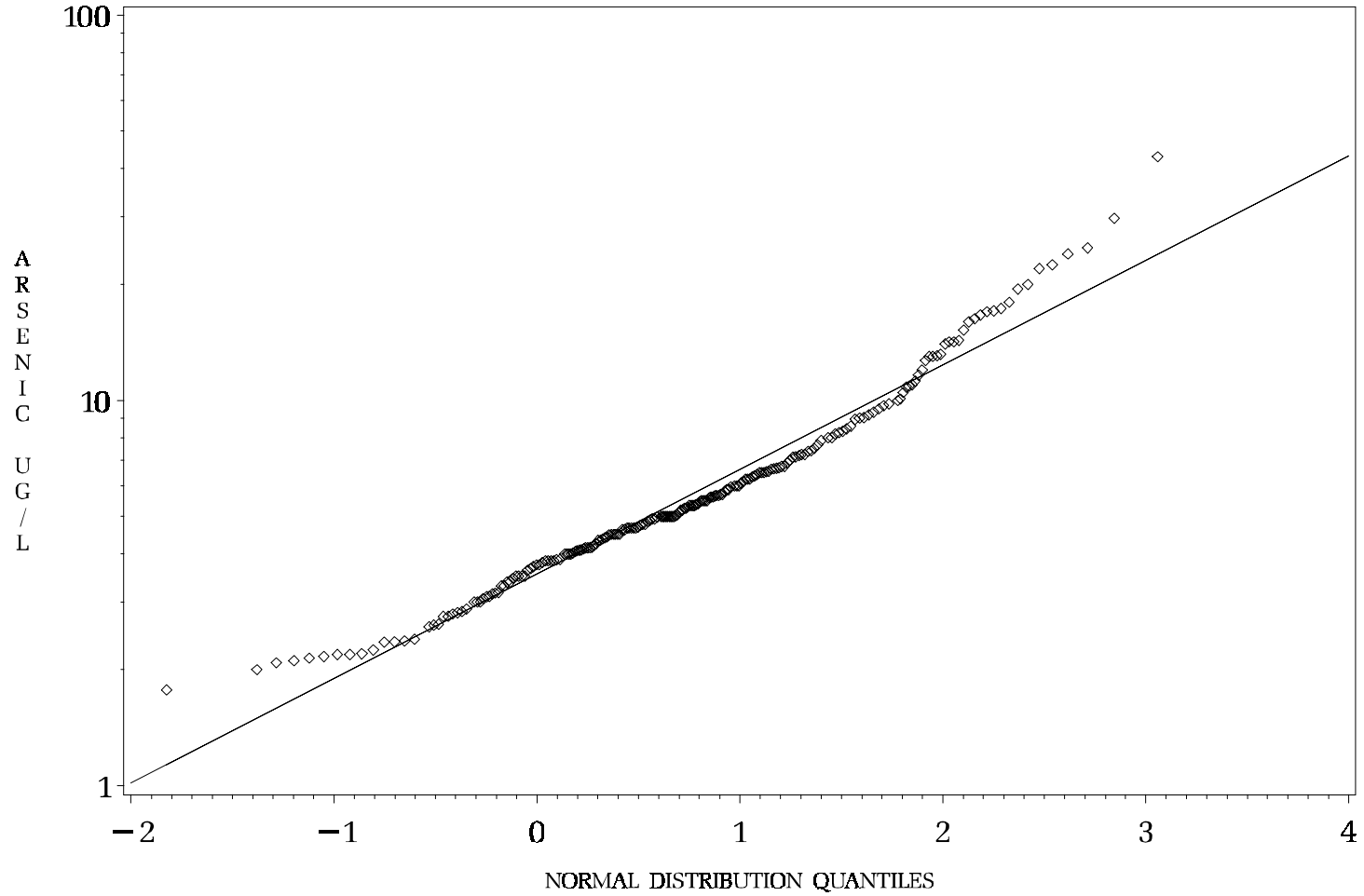


Figure B−21: System means of CWS GW arsenic concentrations for OK, Log−normal probability plot

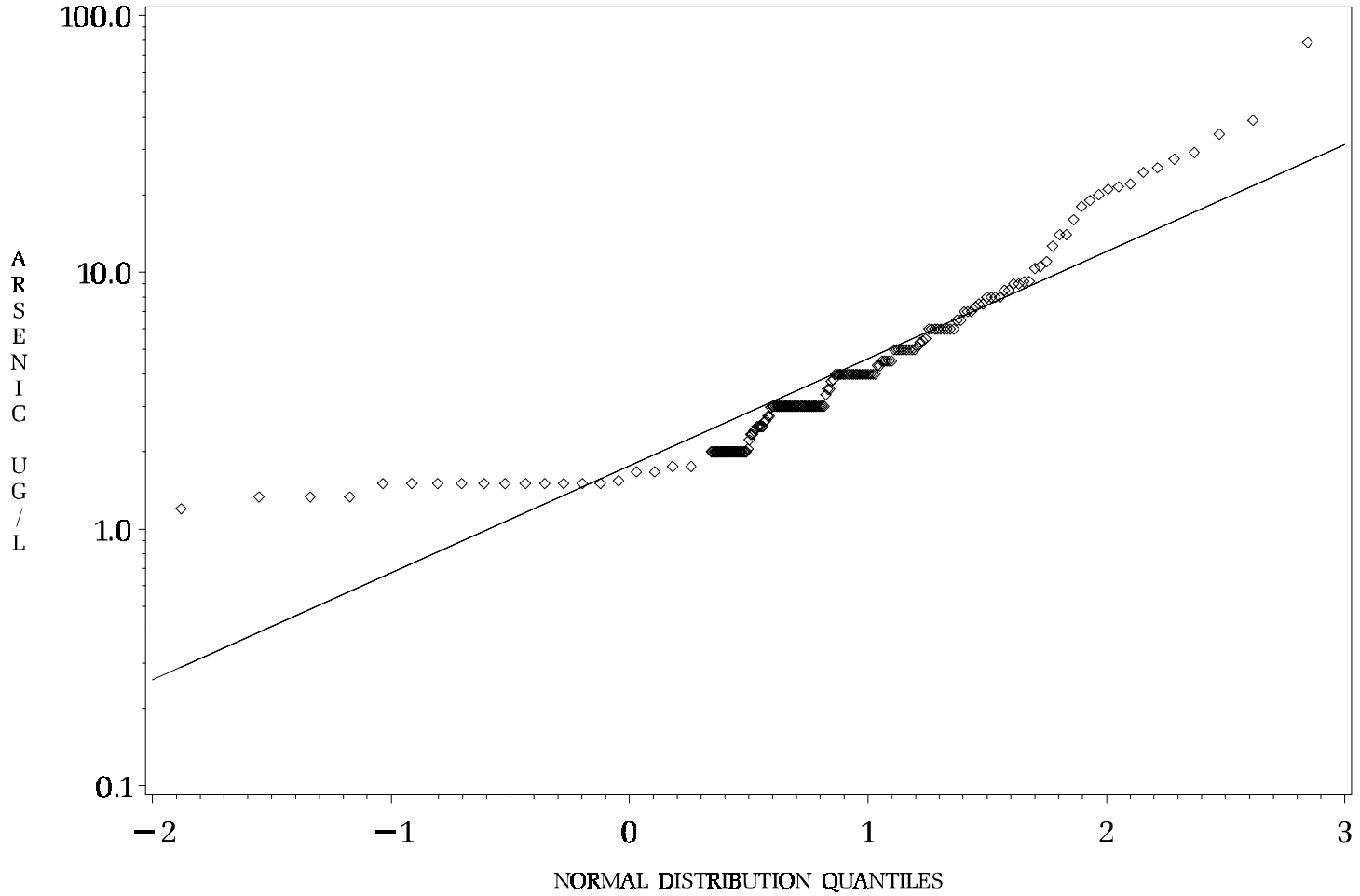


Figure B−22: System means of CWS GW arsenic concentrations for OR, Log−normal probability plot

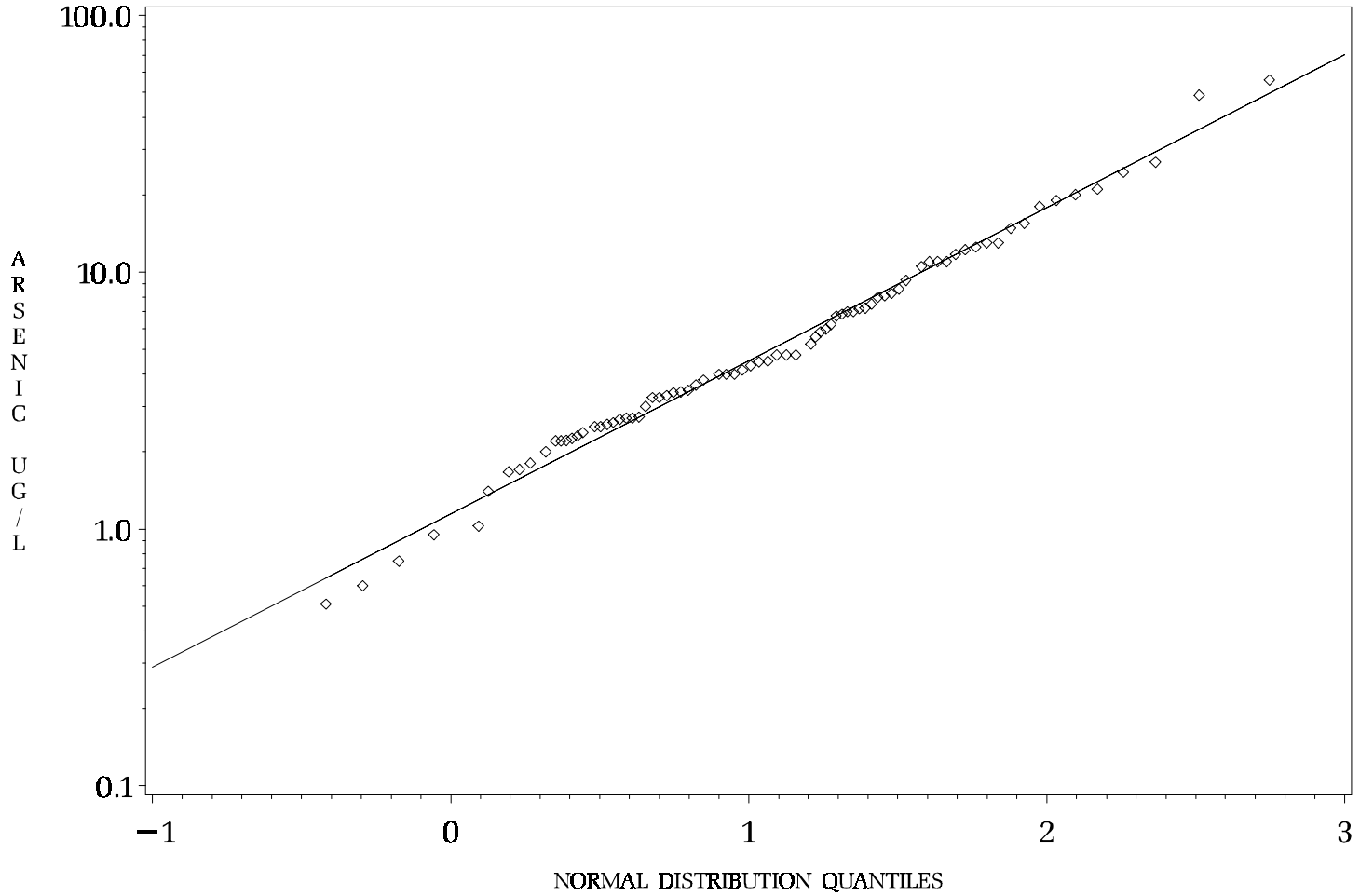


Figure B−23: System means of CWS GW arsenic concentrations for TX, Log−normal probability plot

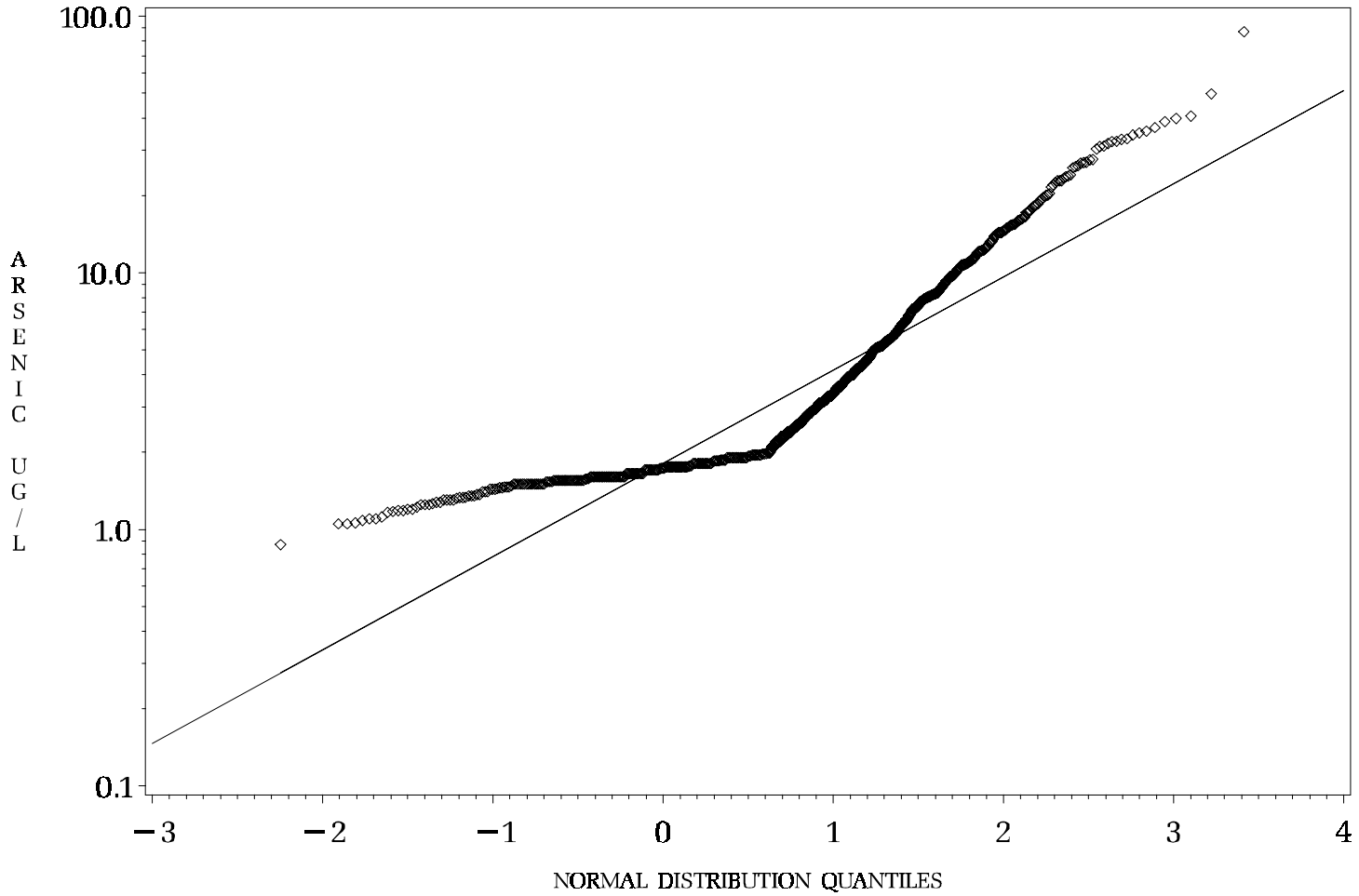


Figure B−24: System means of CWS GW arsenic concentrations for UT, Log−normal probability plot

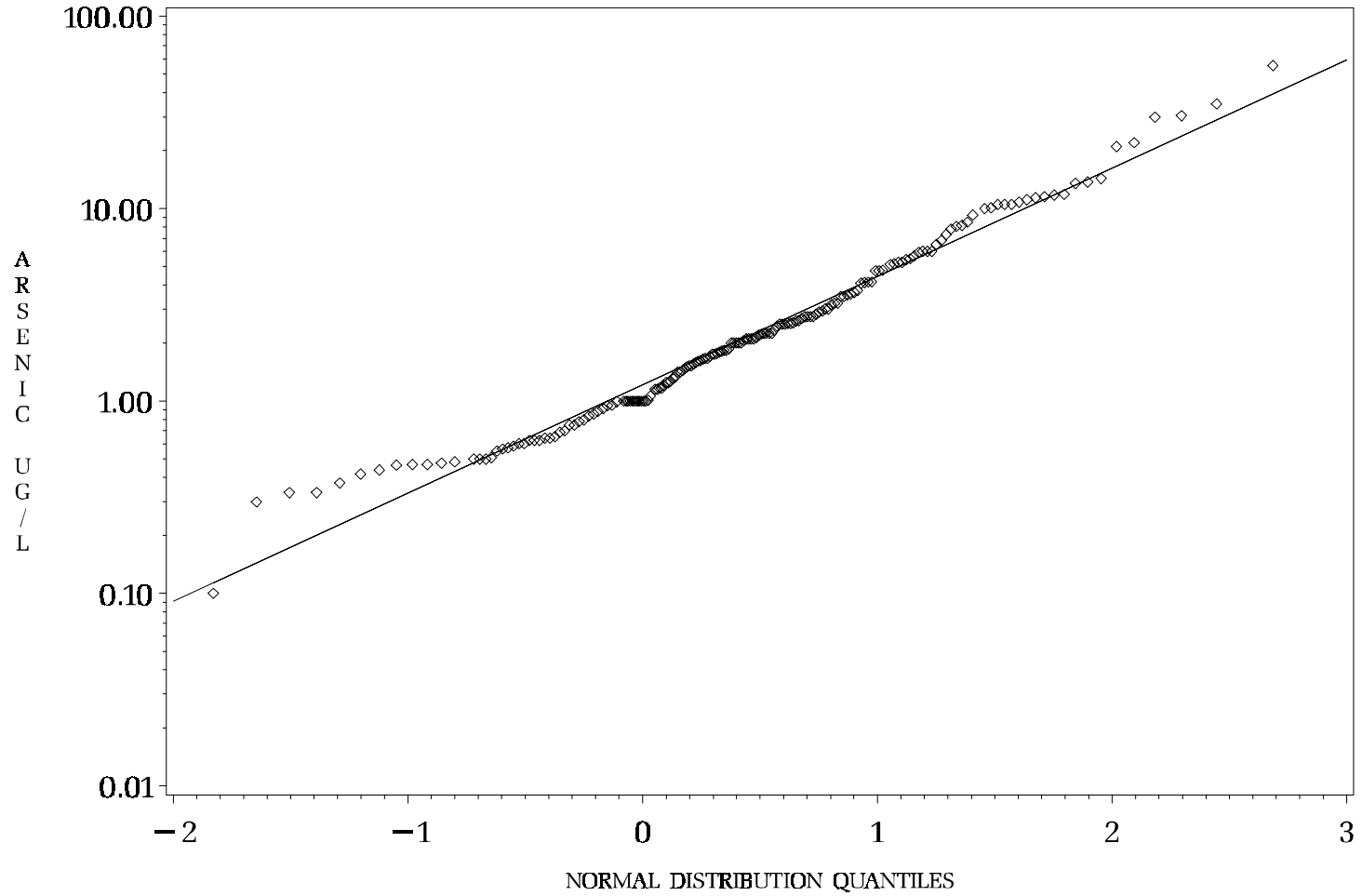


Figure B−25: System means of CWS SW arsenic concentrations for AK, Log−normal probability plot

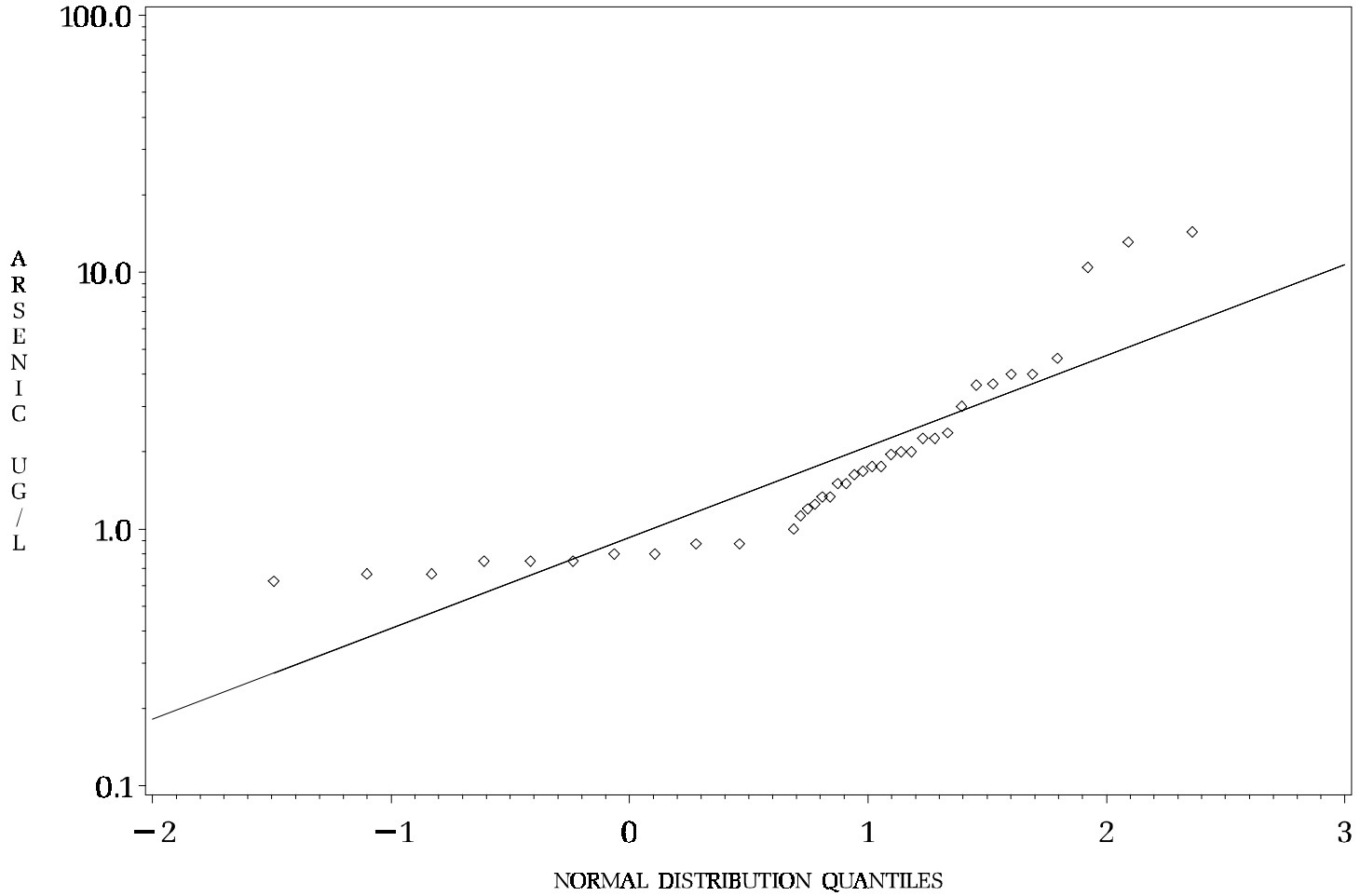


Figure B−26: System means of CWS SW arsenic concentrations for AL, Log−normal probability plot

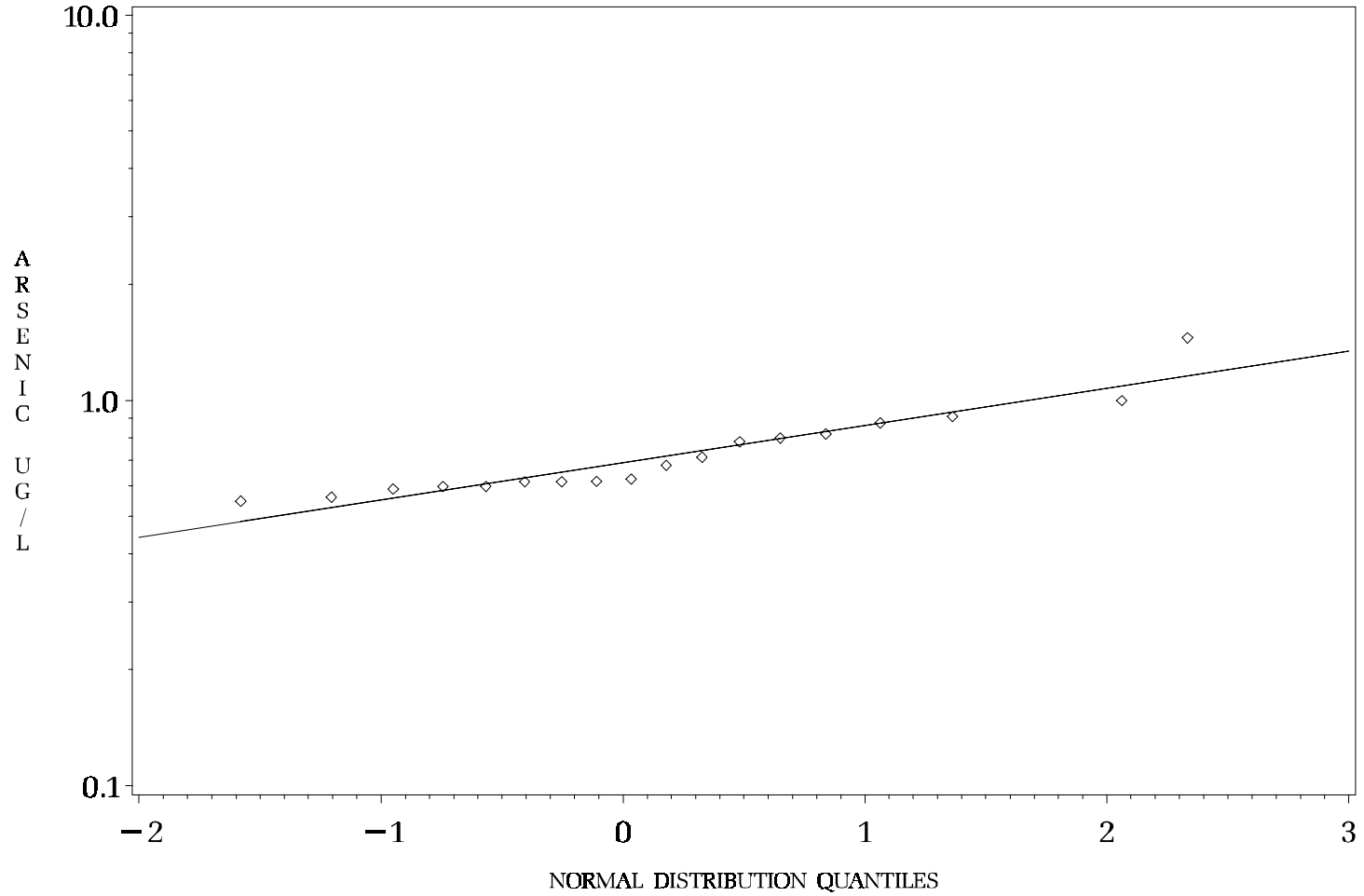


Figure B−27: System means of CWS SW arsenic concentrations for AZ, Log−normal probability plot

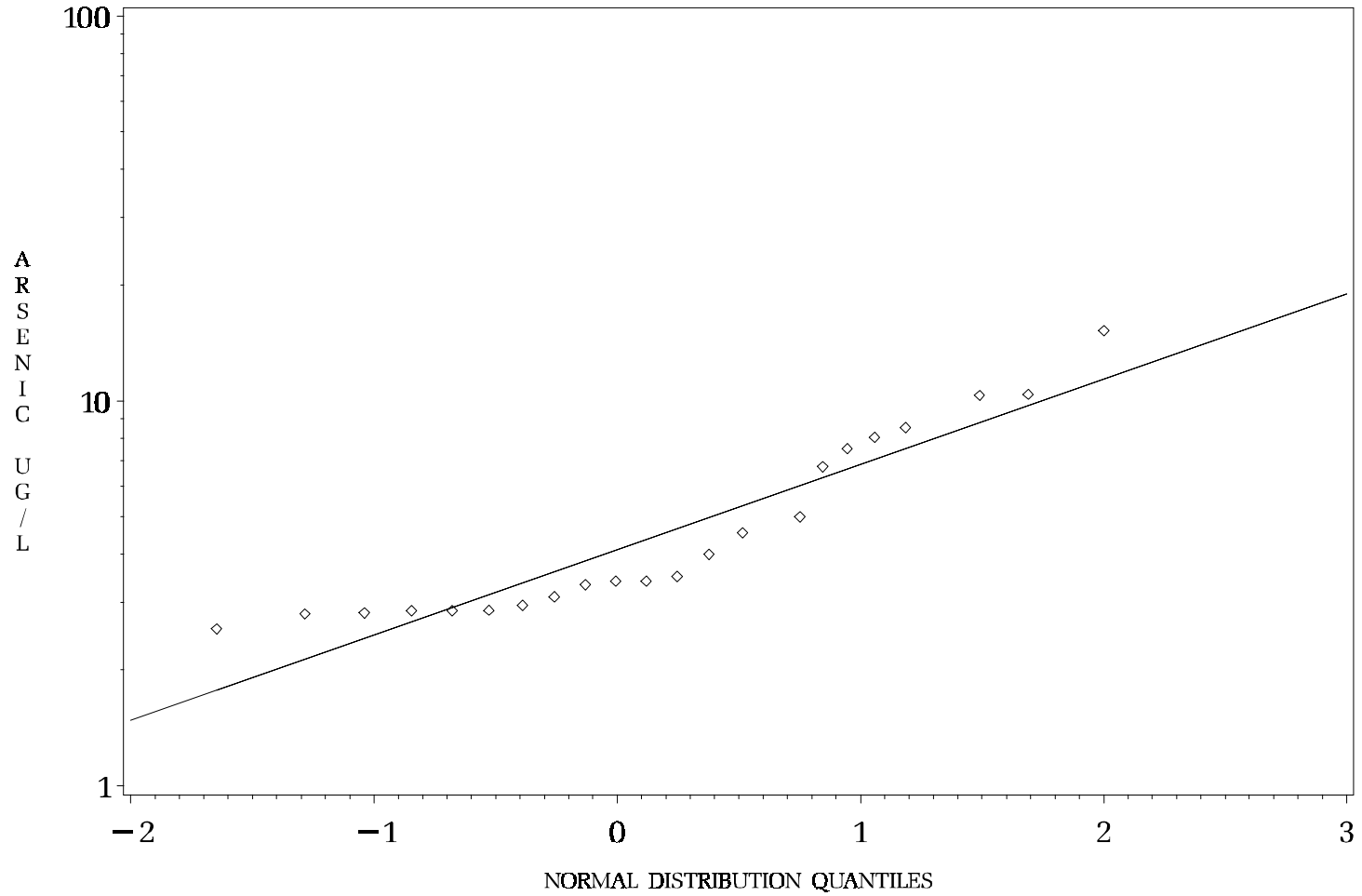


Figure B−28: System means of CWS SW arsenic concentrations for CA, Log−normal probability plot

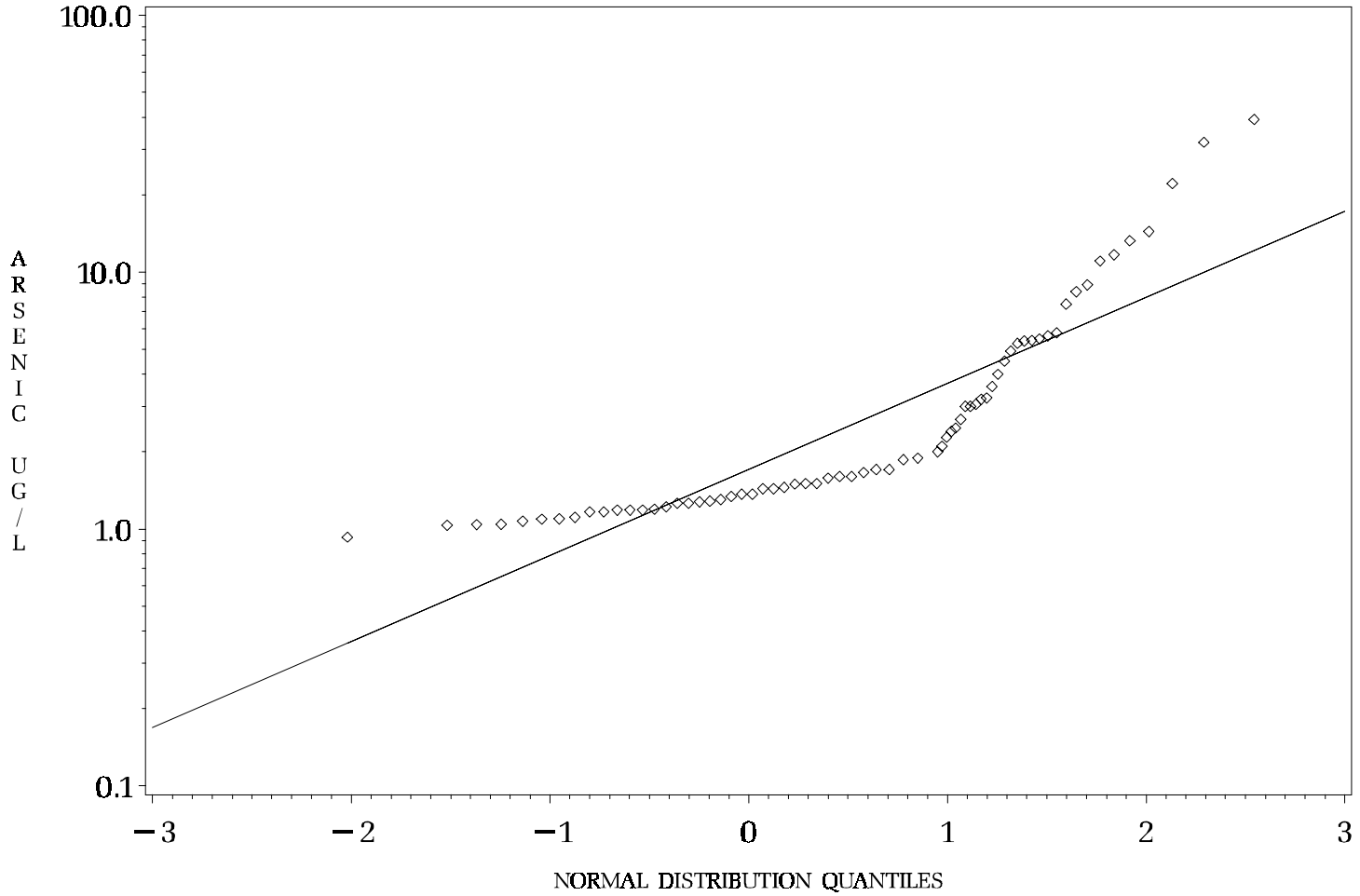


Figure B−29: System means of CWS SW arsenic concentrations for IL, Log−normal probability plot

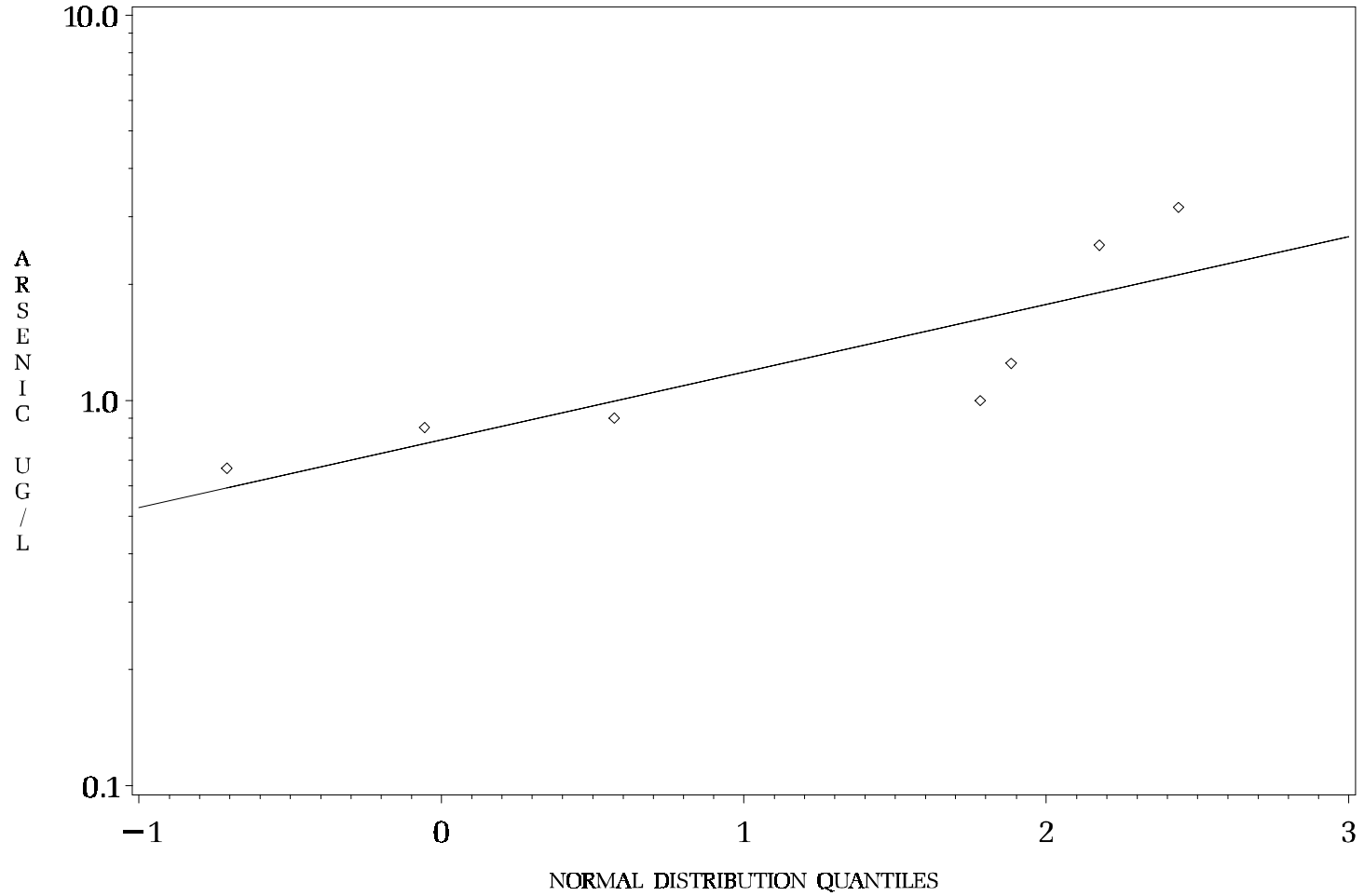


Figure B−30: System means of CWS SW arsenic concentrations for KS, Log−normal probability plot

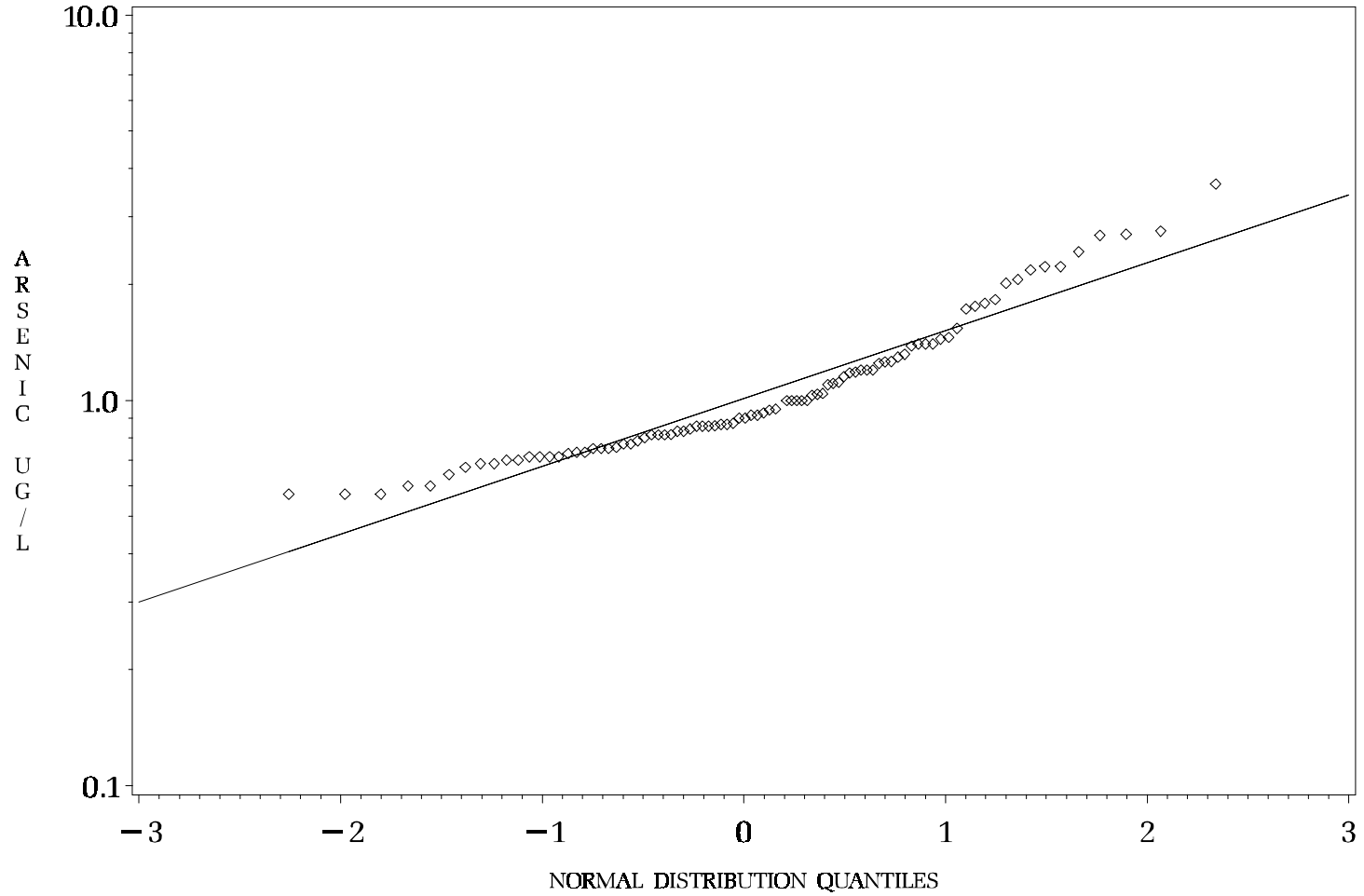


Figure B−31: System means of CWS SW arsenic concentrations for KY, Log−normal probability plot

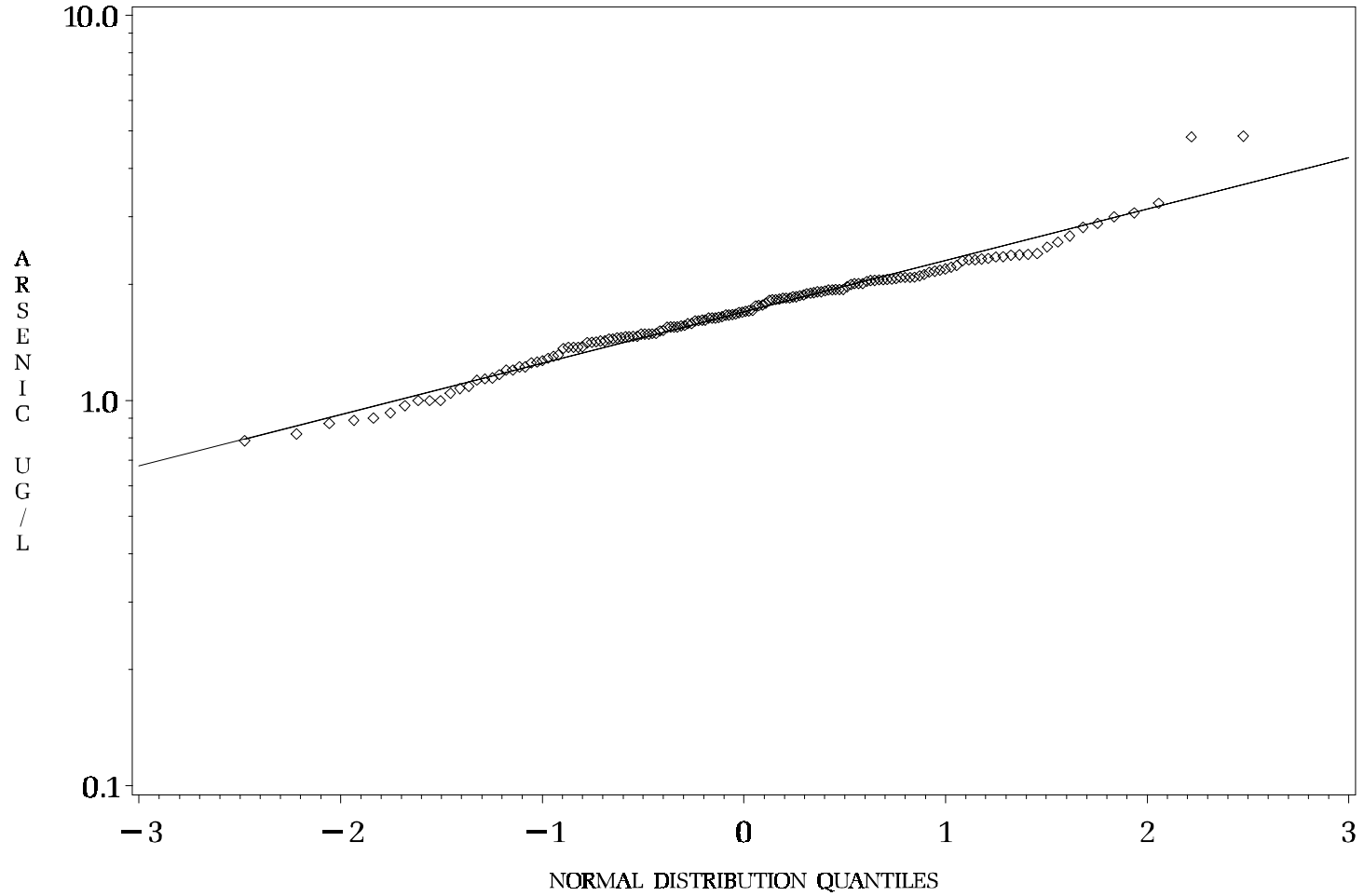


Figure B−32: System means of CWS SW arsenic concentrations for ME, Log−normal probability plot

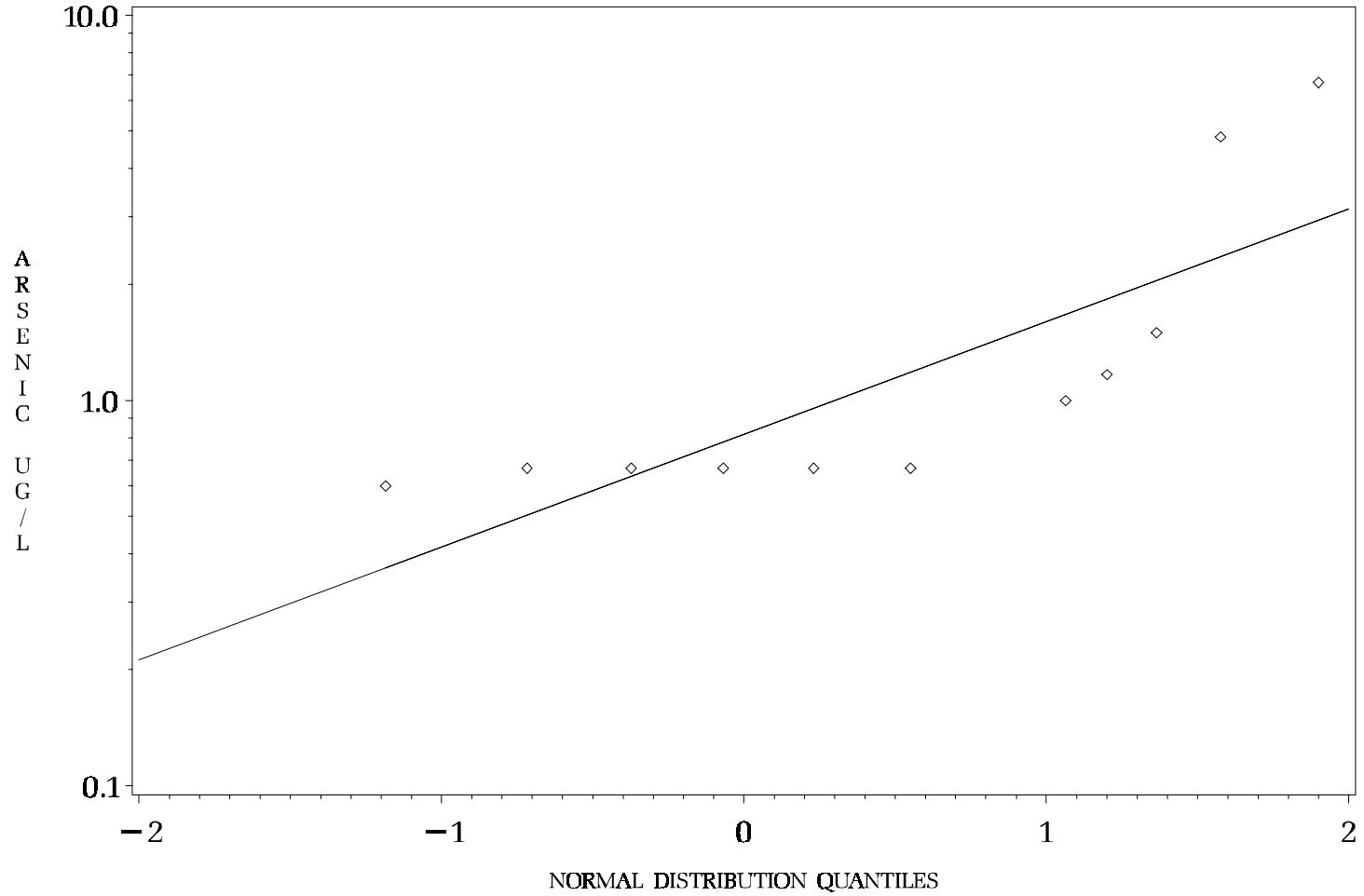


Figure B−33: System means of CWS SW arsenic concentrations for MI, Log−normal probability plot

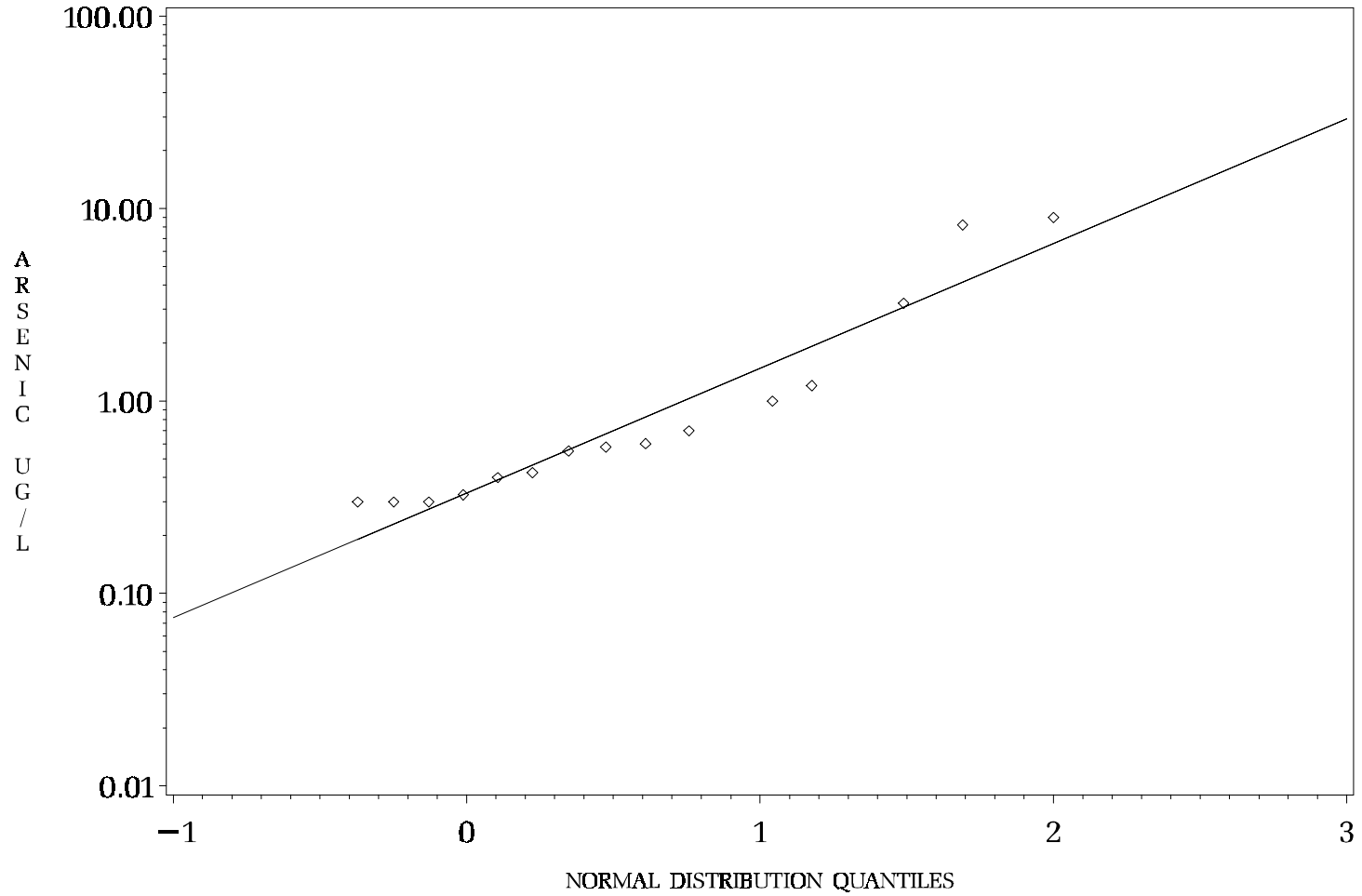


Figure B−34: System means of CWS SW arsenic concentrations for MN, Log−normal probability plot

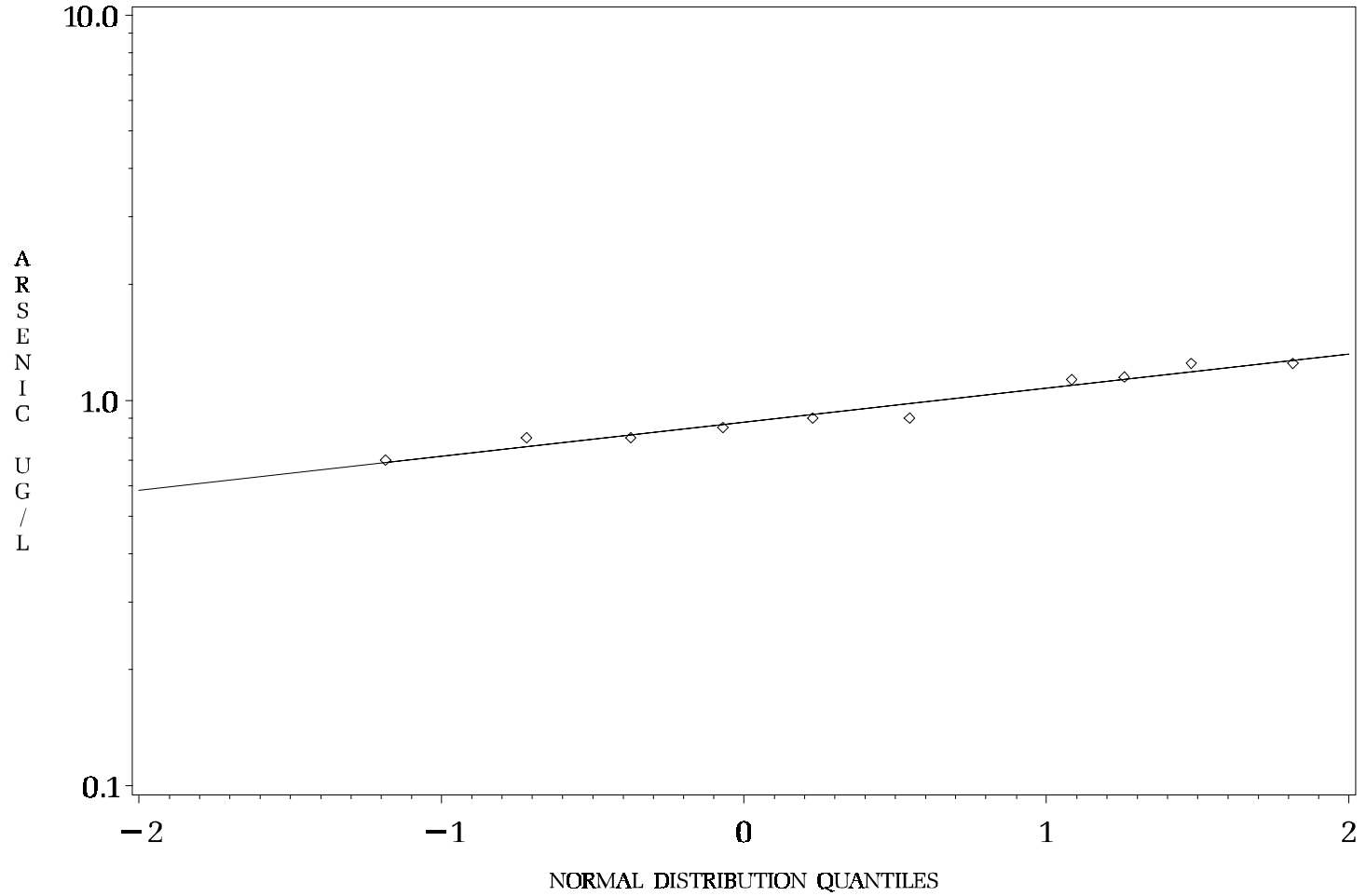


Figure B−35: System means of CWS SW arsenic concentrations for MO, Log−normal probability plot

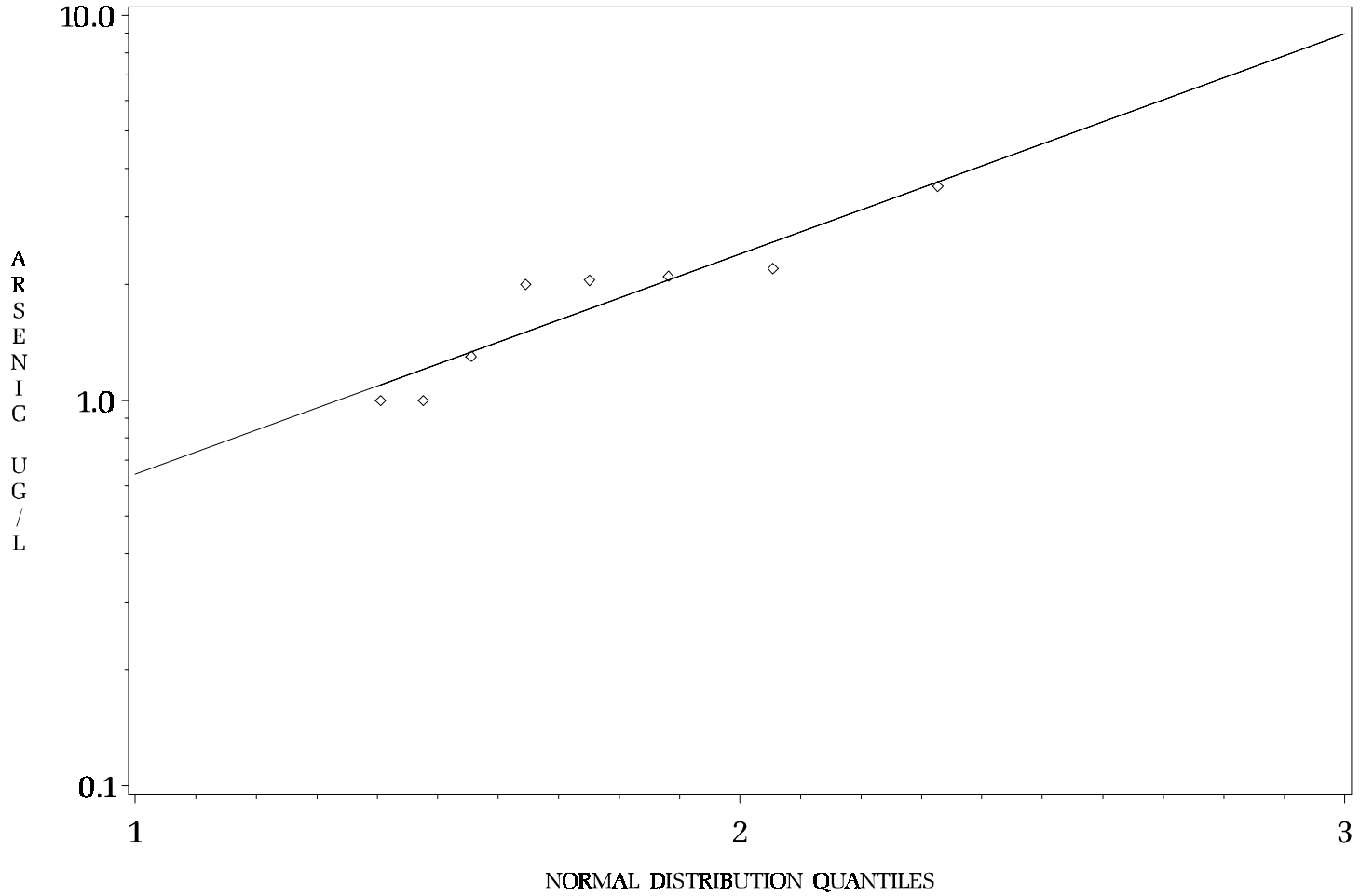


Figure B−36: System means of CWS SW arsenic concentrations for MT, Log−normal probability plot

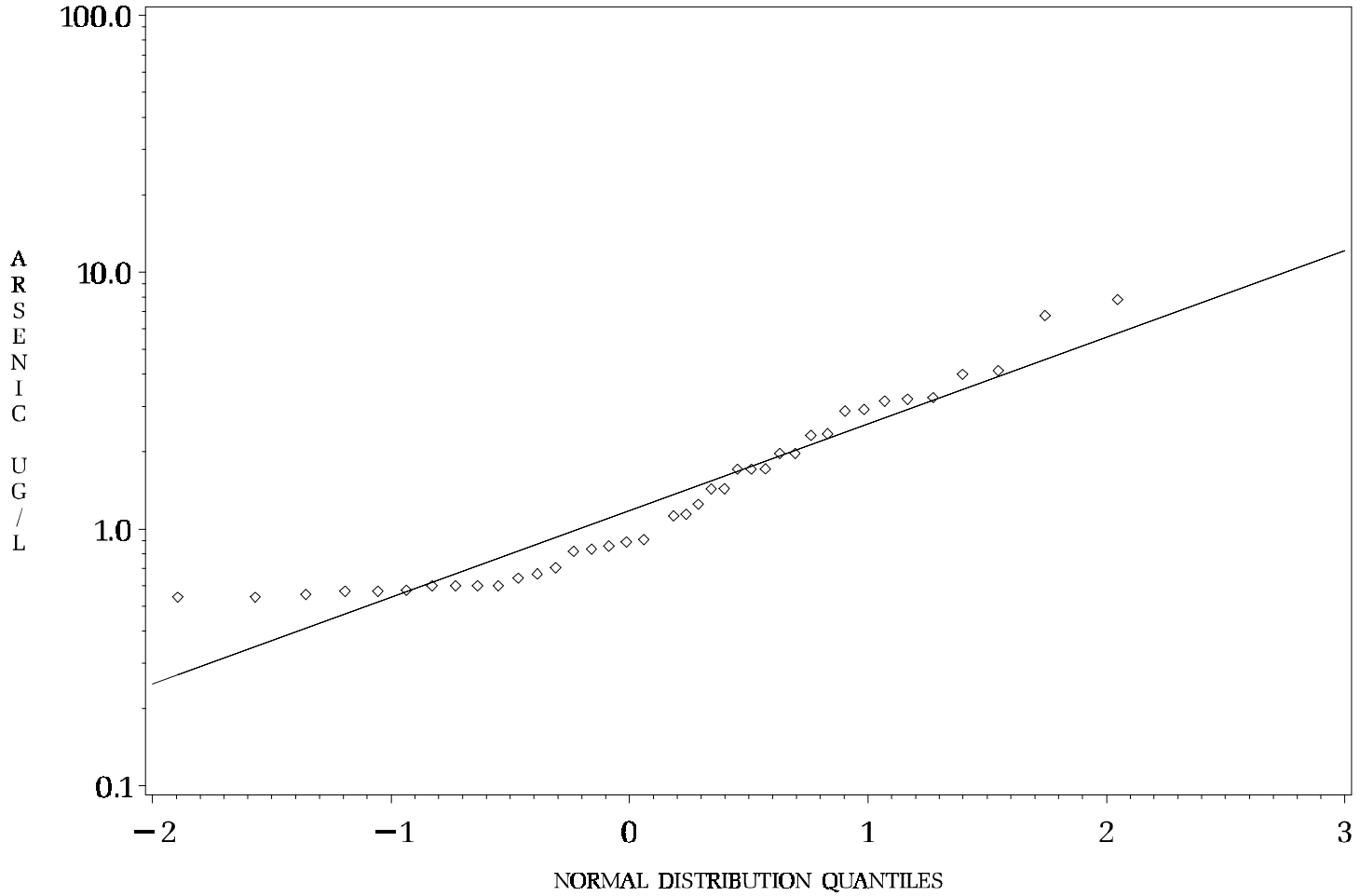


Figure B-37: System means of CWS SW arsenic concentrations for NC, Log-normal probability plot

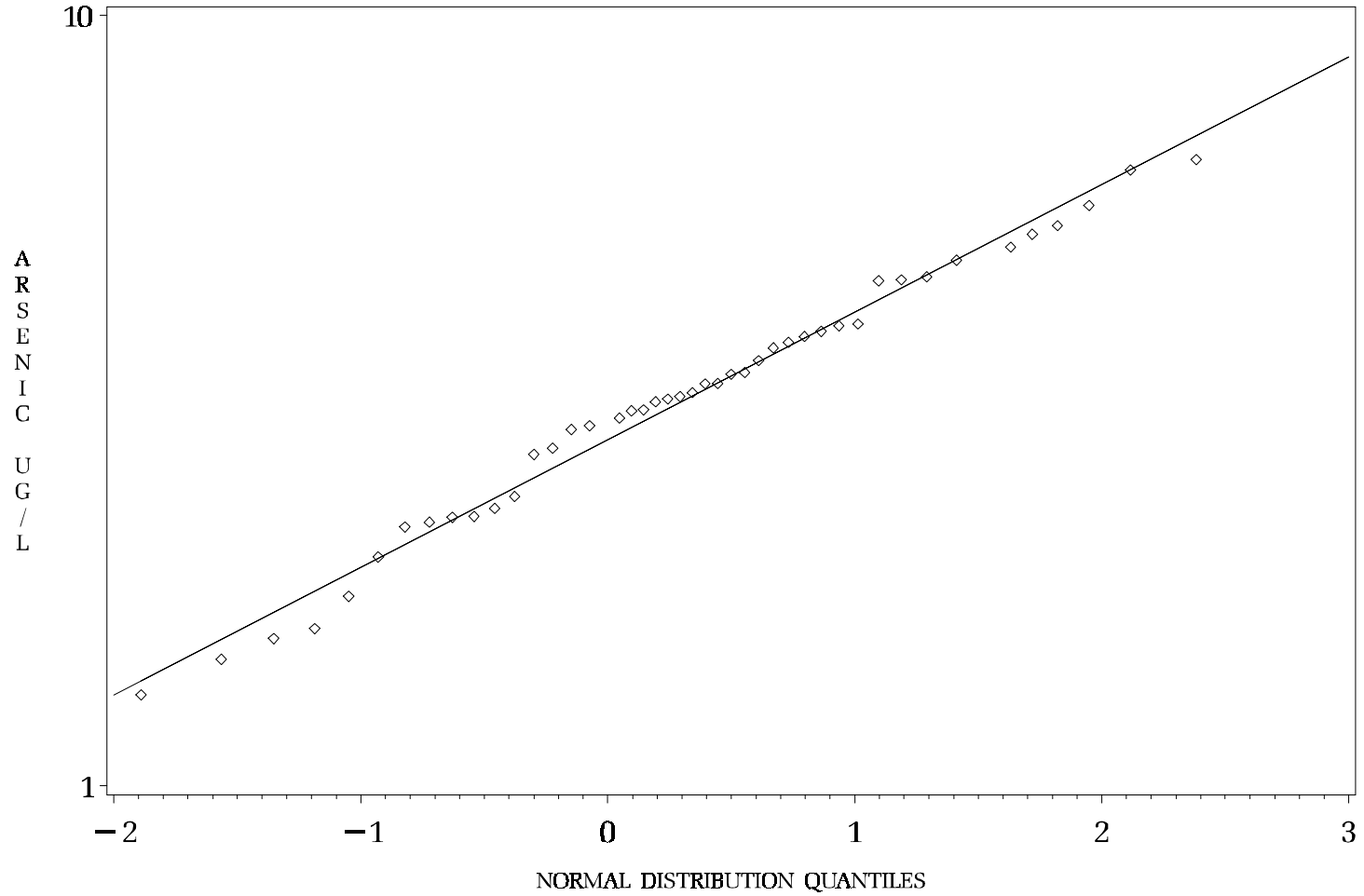


Figure B−38: System means of CWS SW arsenic concentrations for ND, Log−normal probability plot

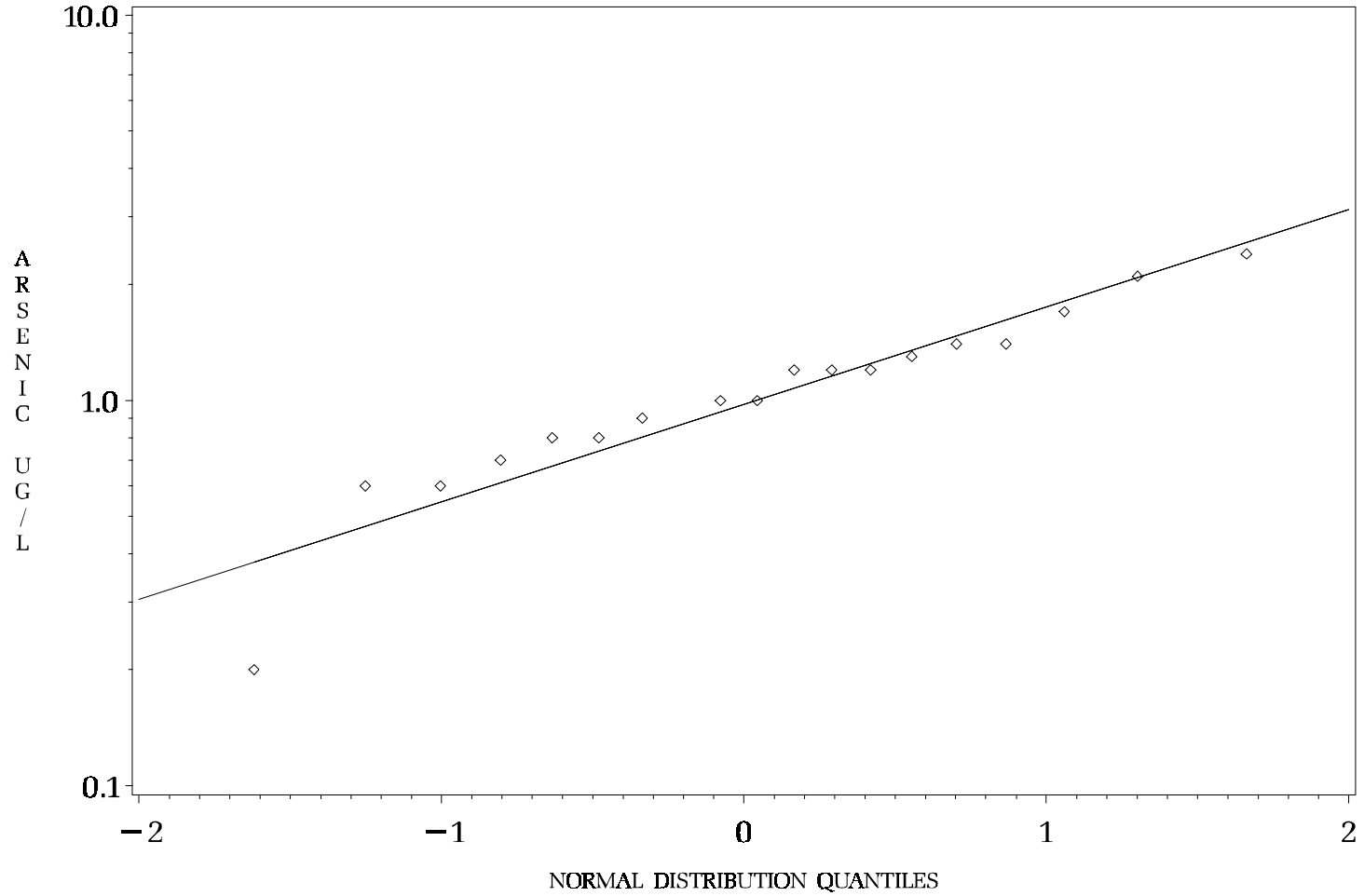


Figure B−39: System means of CWS SW arsenic concentrations for NH, Log−normal probability plot

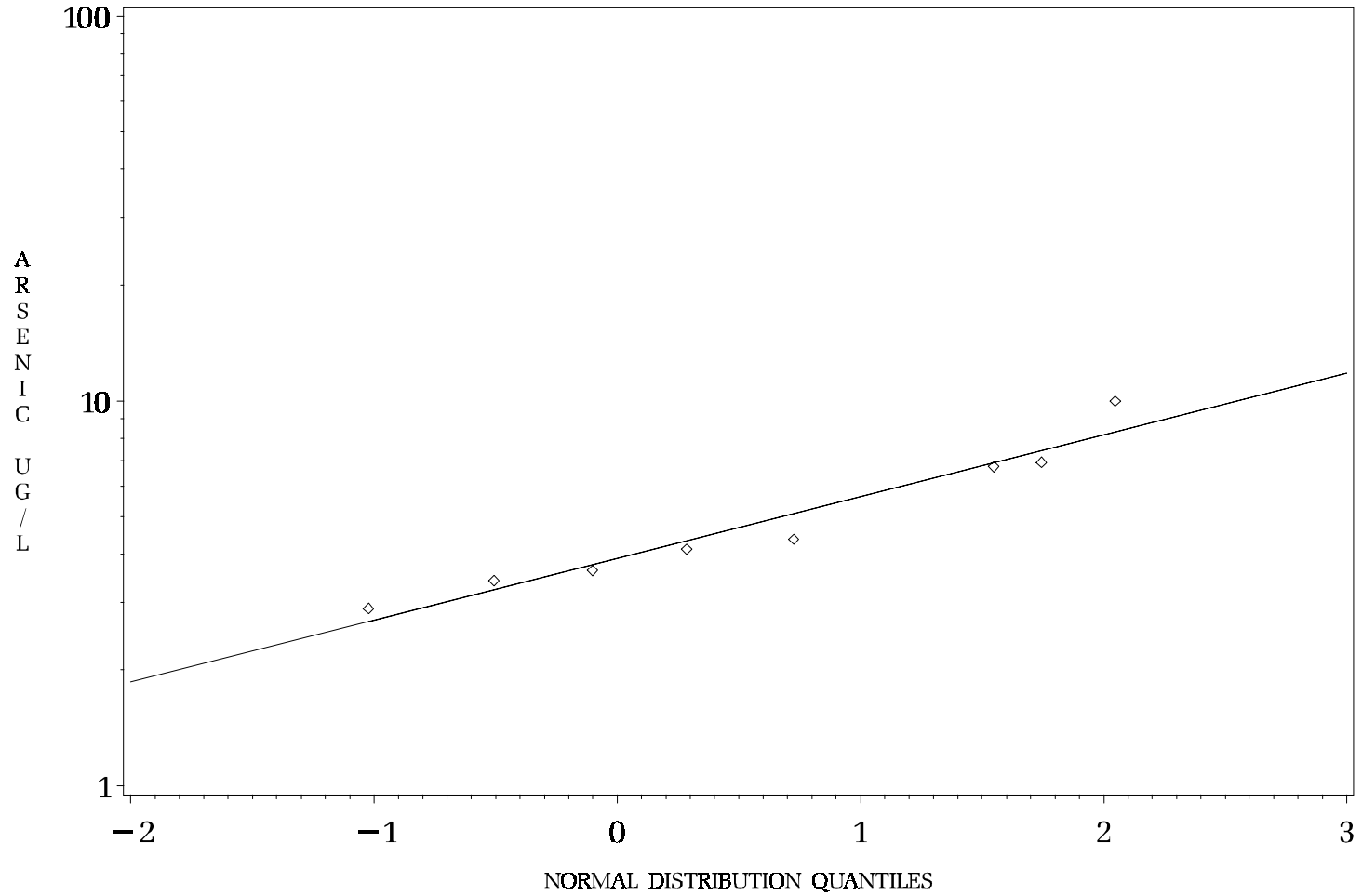


Figure B−40: System means of CWS SW arsenic concentrations for NJ, Log−normal probability plot

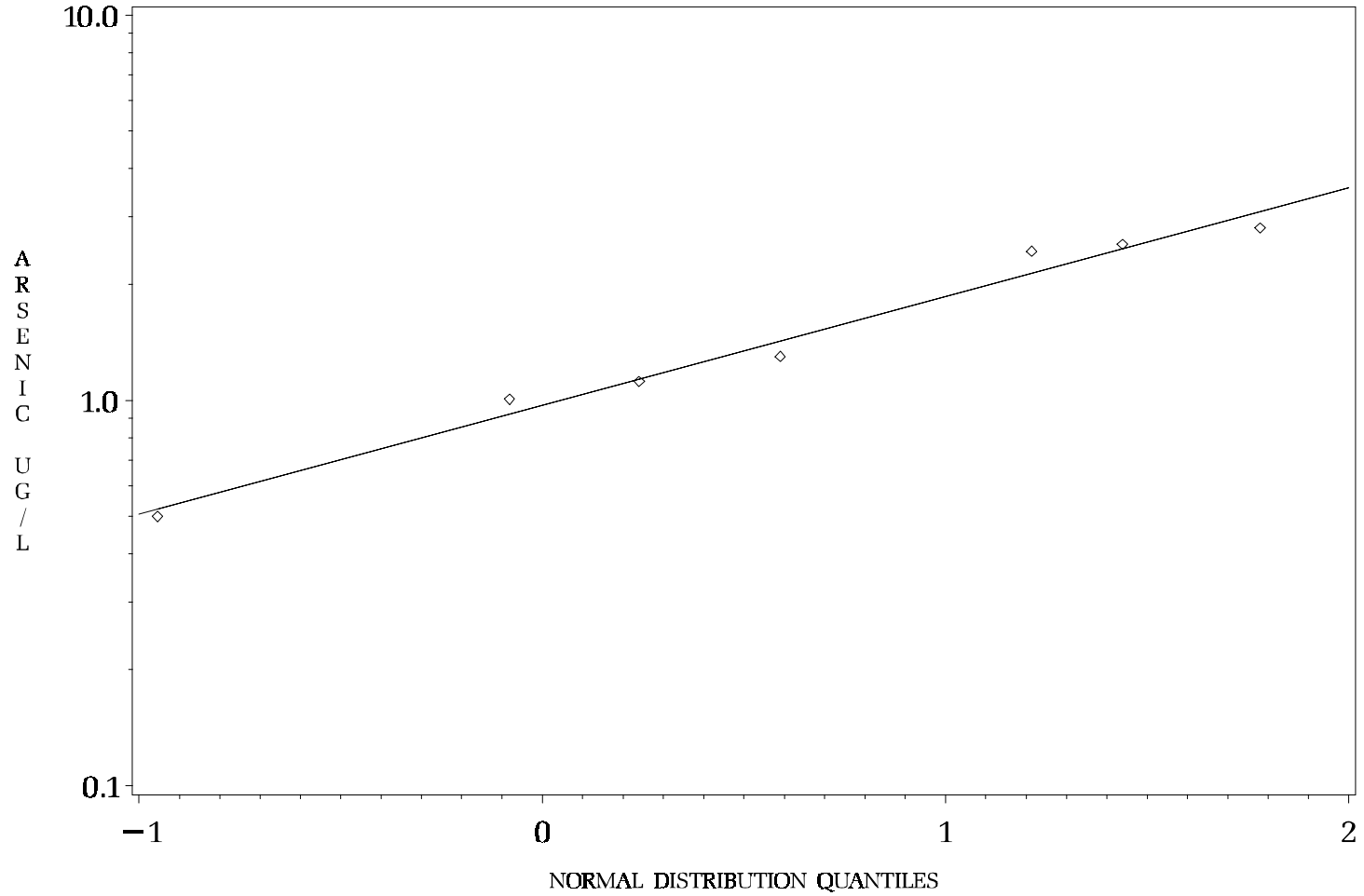


Figure B−41: System means of CWS SW arsenic concentrations for NM, Log−normal probability plot

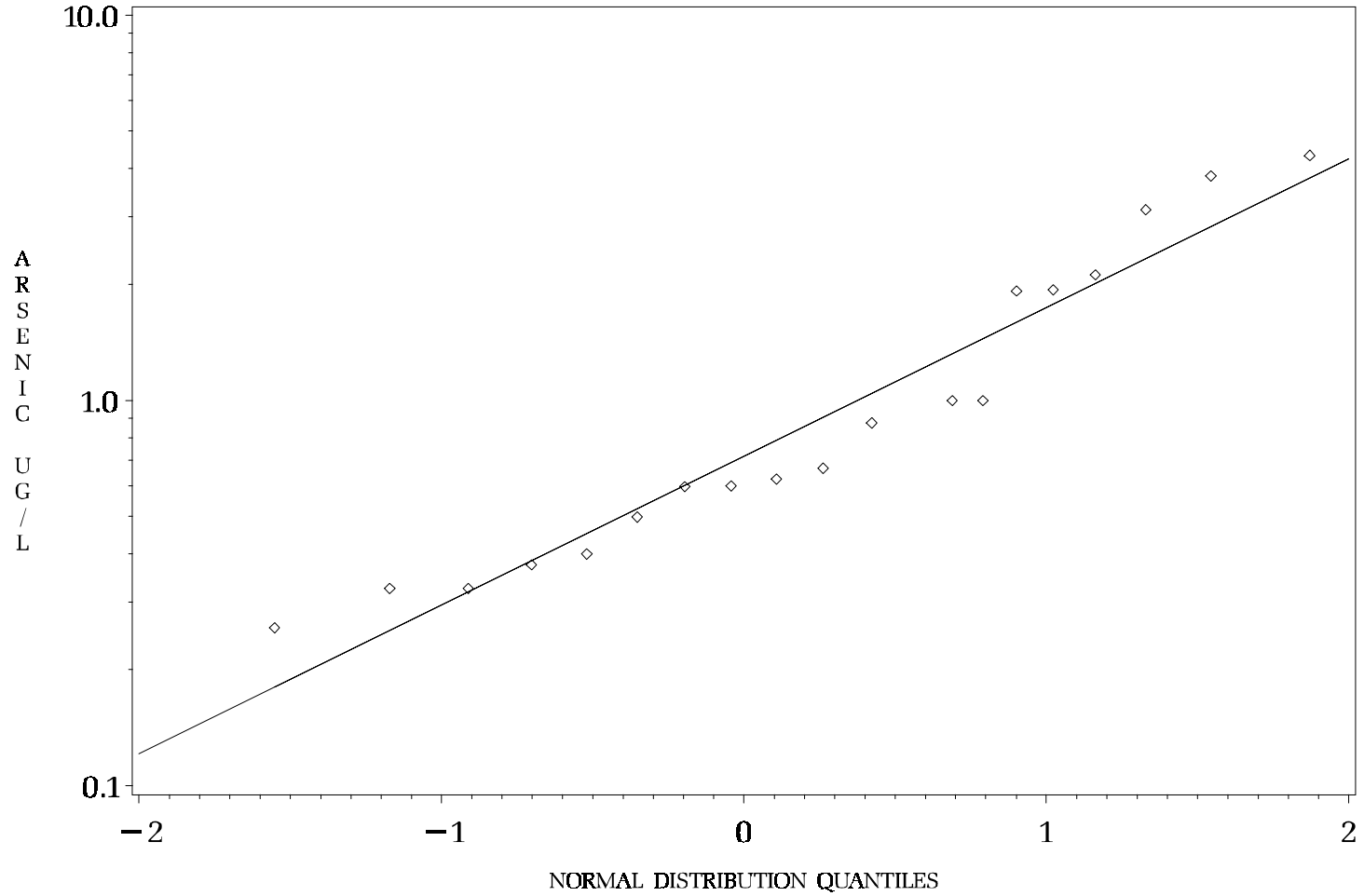


Figure B−42: System means of CWS SW arsenic concentrations for NV, Log−normal probability plot

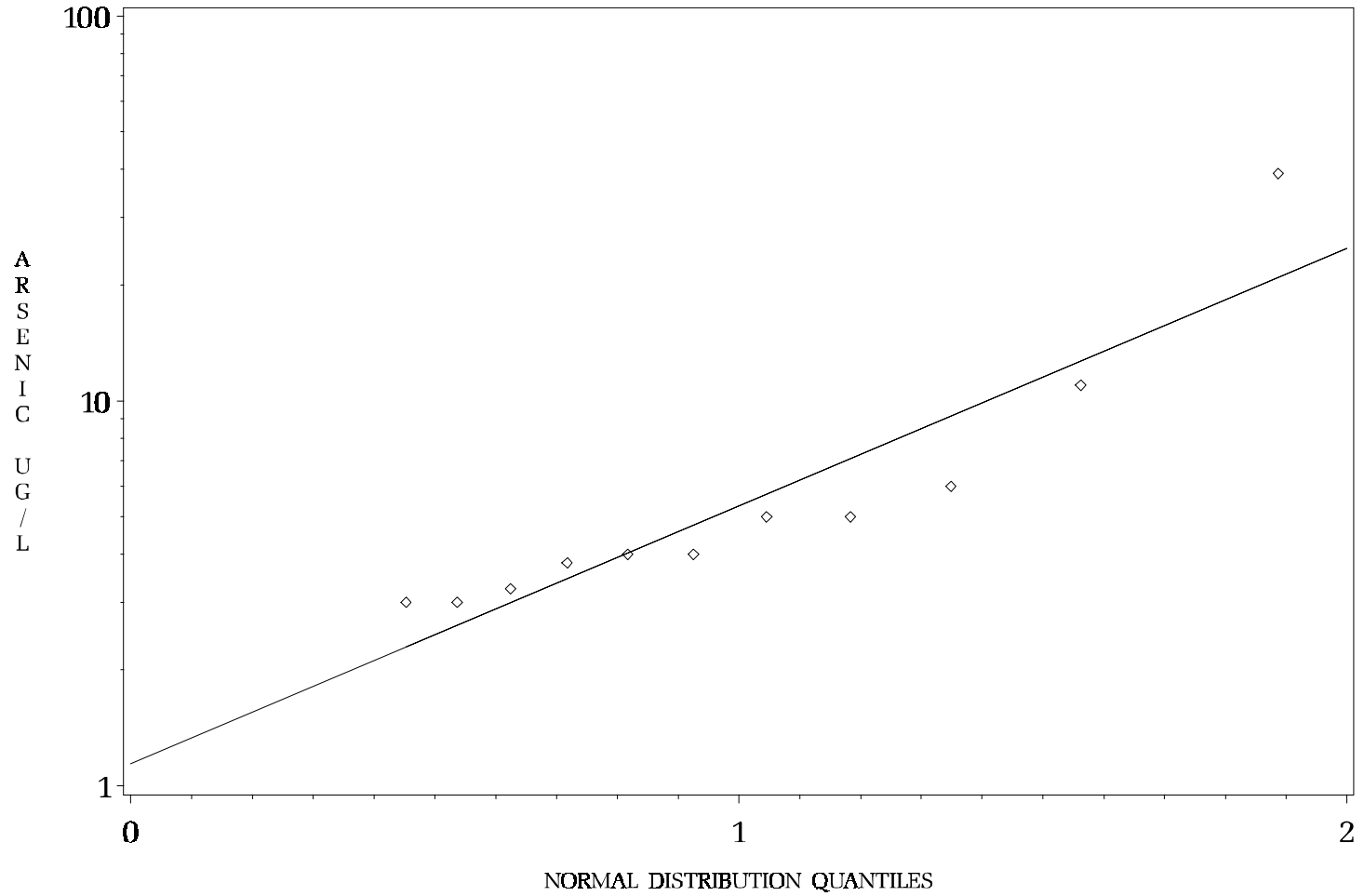


Figure B−43: System means of CWS SW arsenic concentrations for OH, Log−normal probability plot

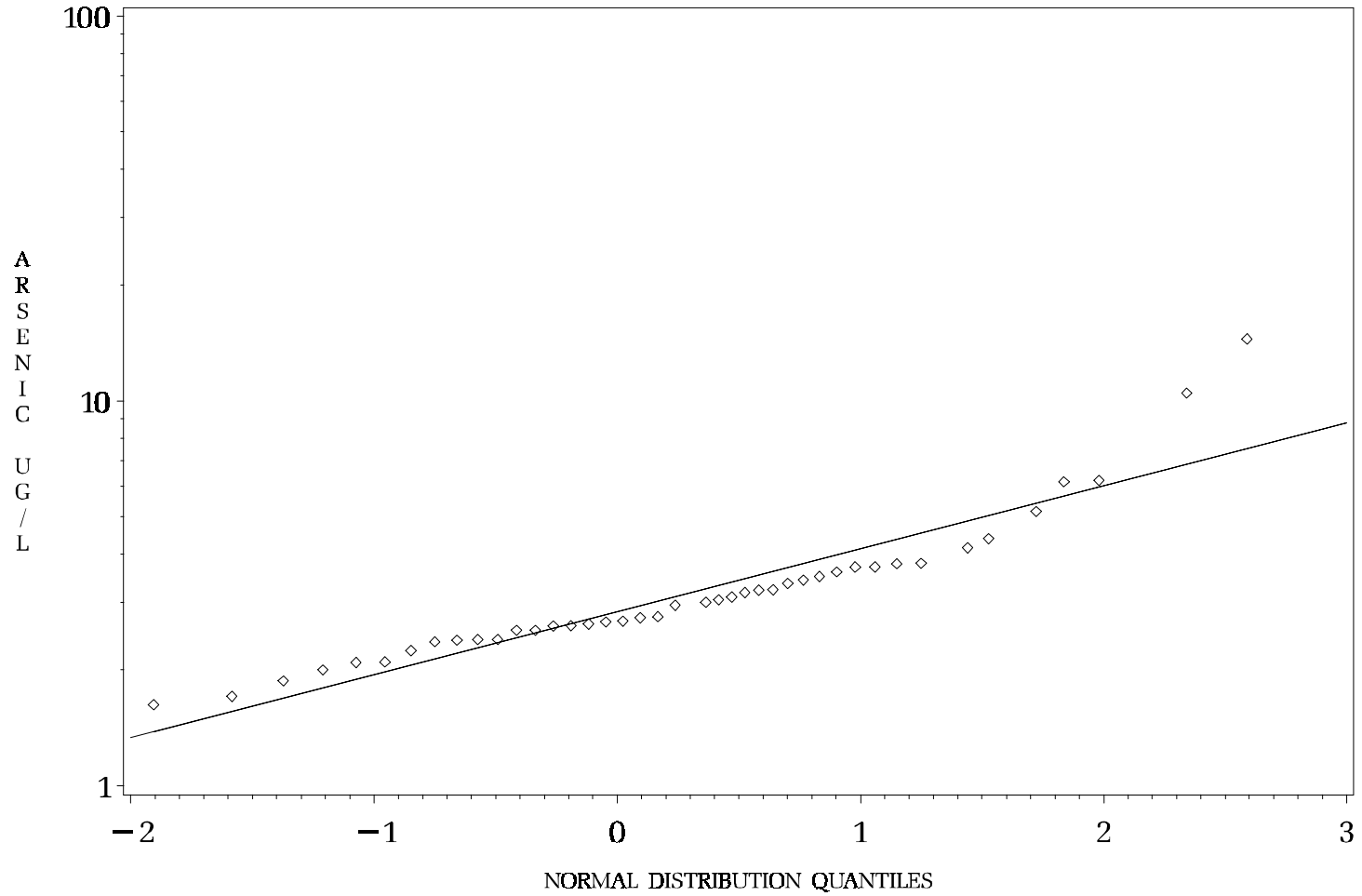


Figure B−44: System means of CWS SW arsenic concentrations for OK, Log−normal probability plot

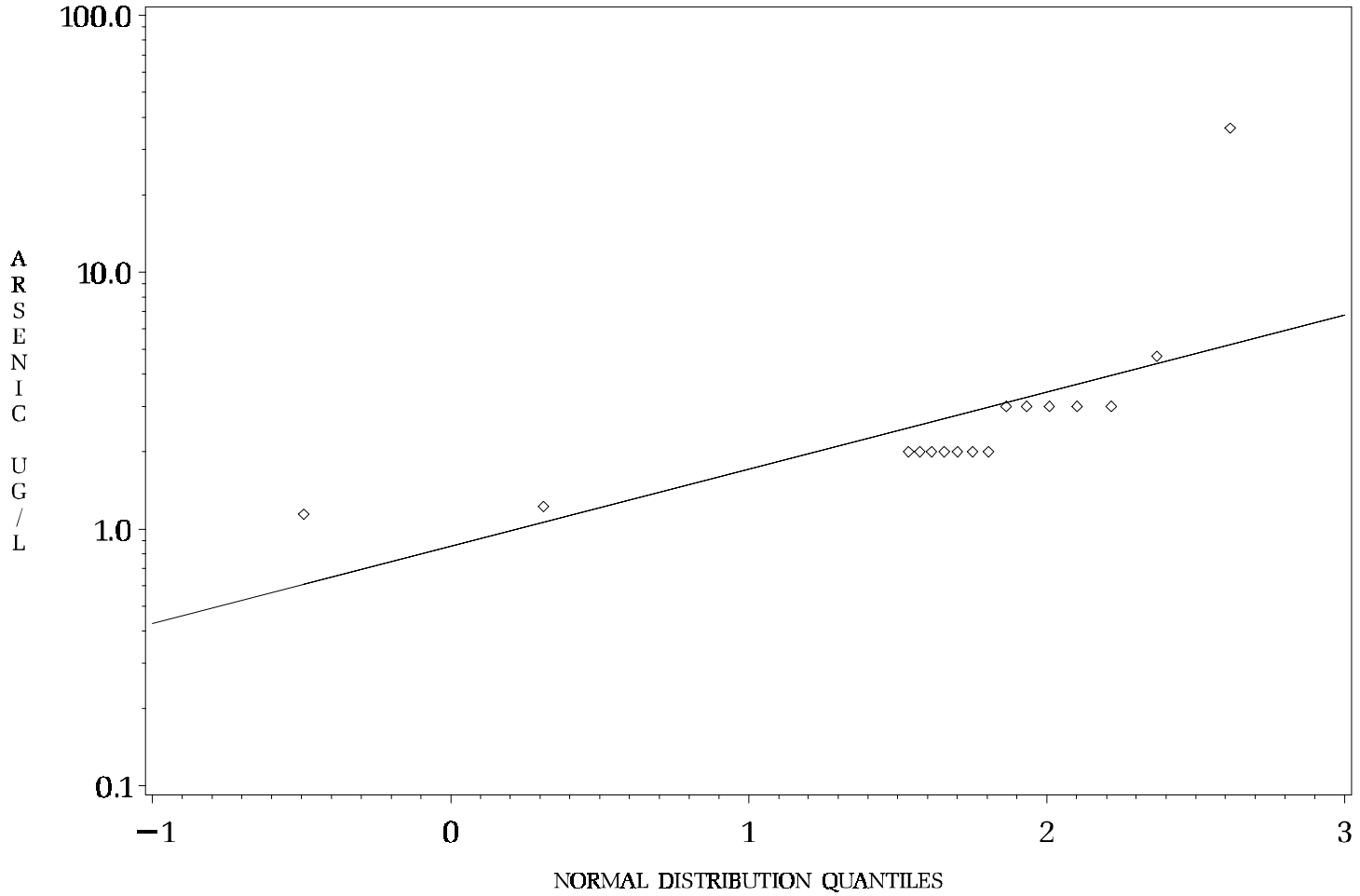


Figure B−45: System means of CWS SW arsenic concentrations for OR, Log−normal probability plot

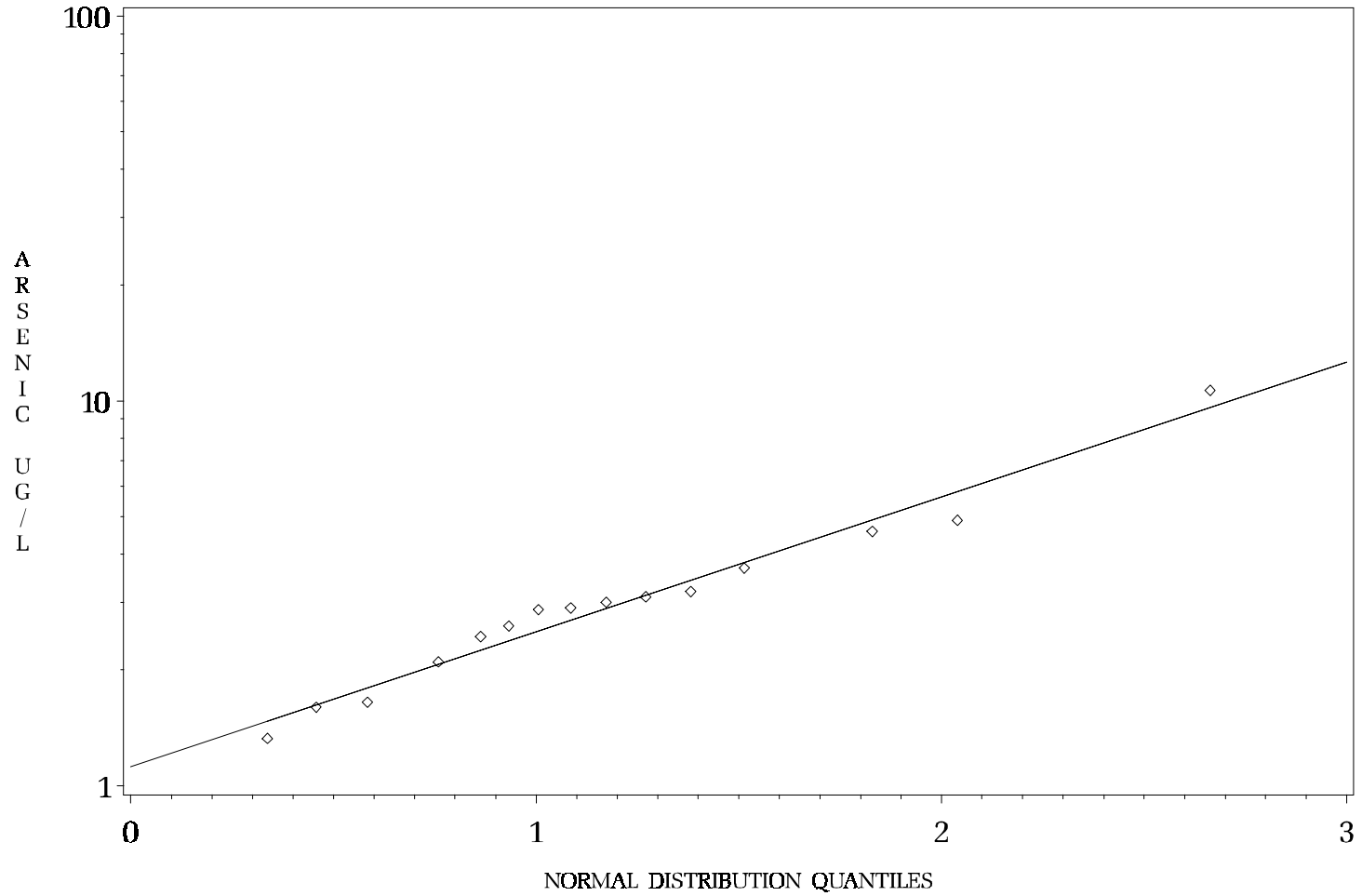


Figure B−46: System means of CWS SW arsenic concentrations for TX, Log−normal probability plot

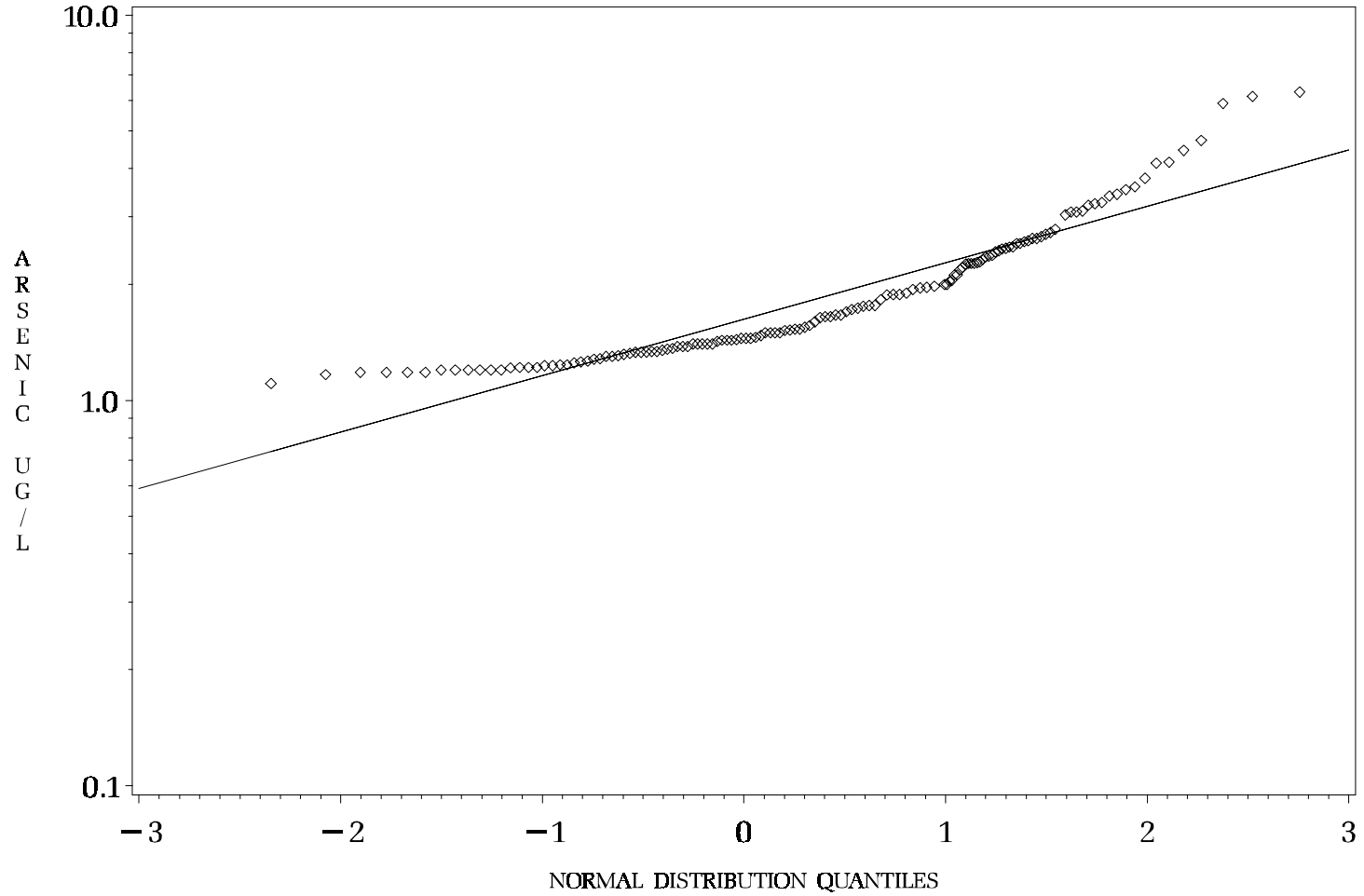


Figure B−47: System means of CWS SW arsenic concentrations for UT, Log−normal probability plot

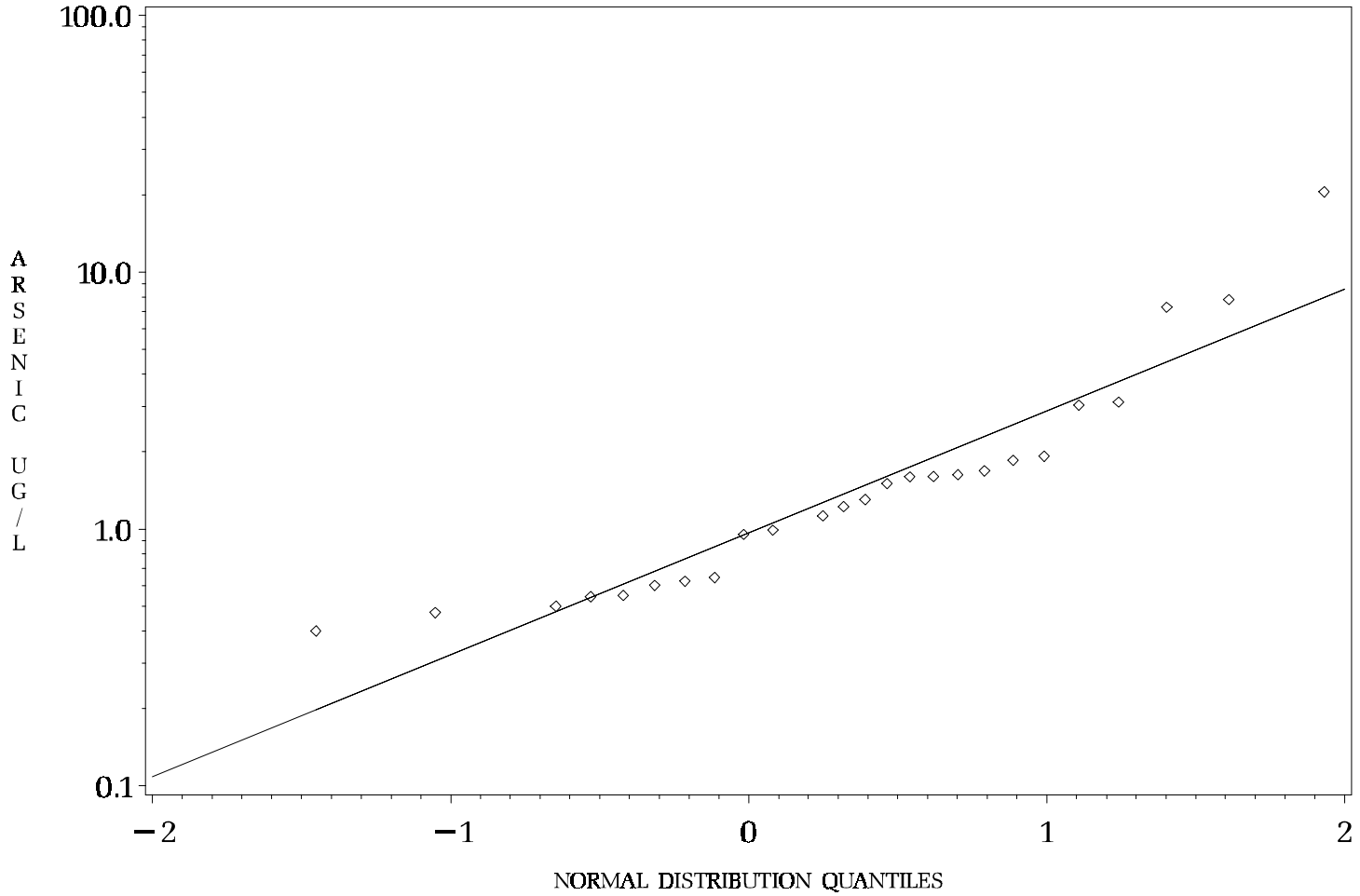


Figure B−48: System means of NTNCWS GW arsenic concentrations for AK, Log−normal probability plot

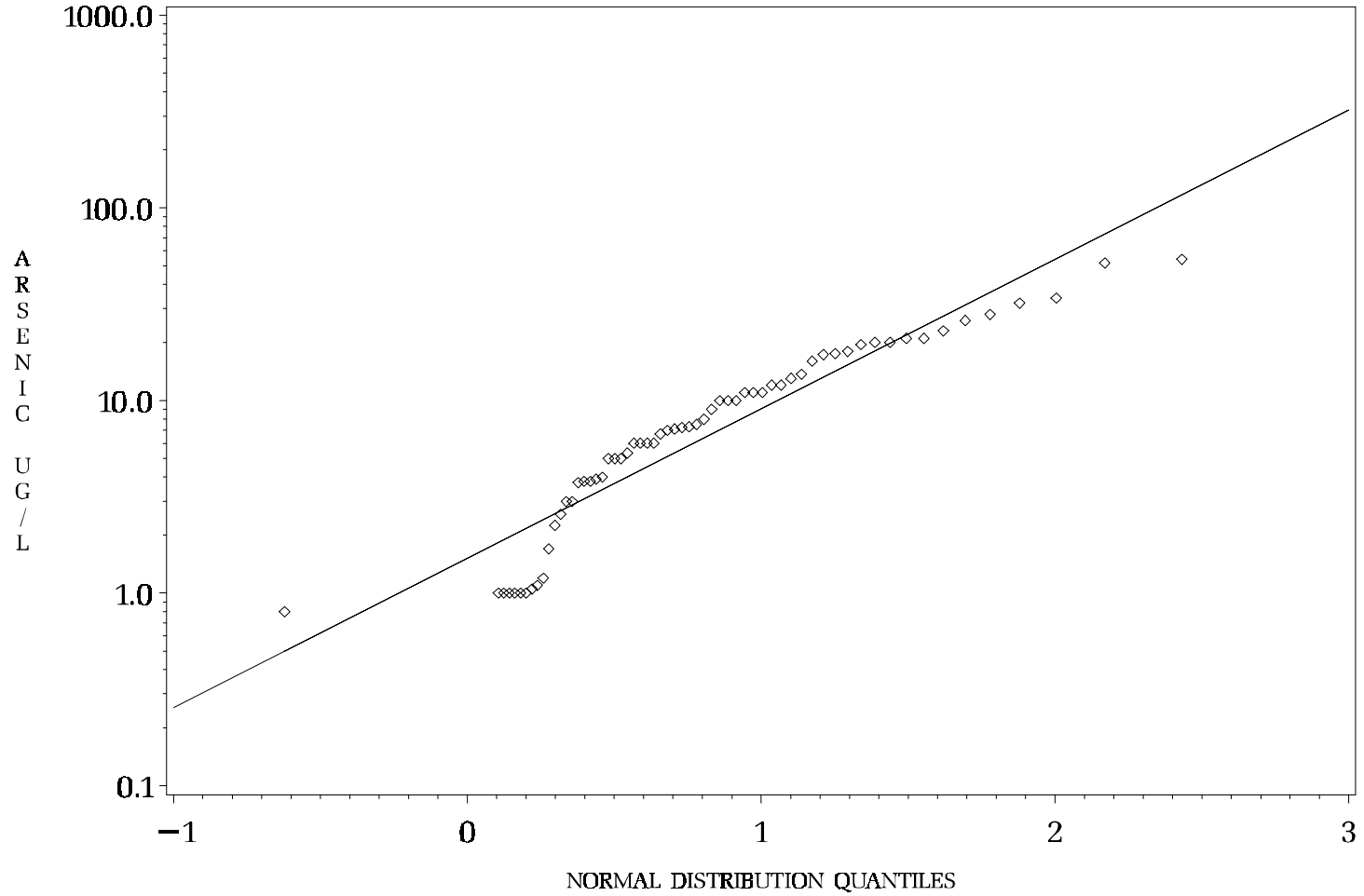


Figure B−49: System means of NTNCWS GW arsenic concentrations for AZ, Log−normal probability plot

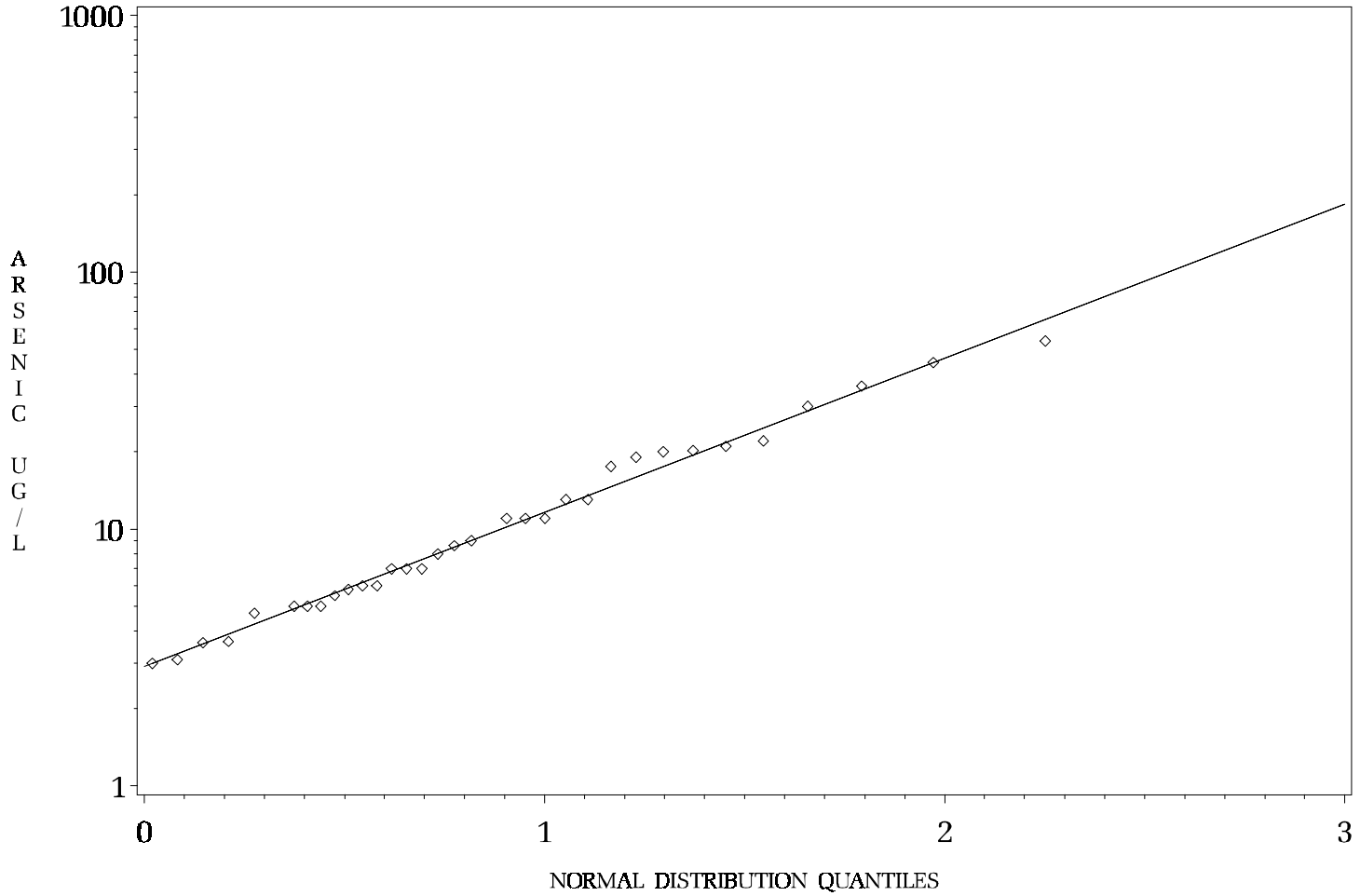


Figure B−50: System means of NTNCWS GW arsenic concentrations for CA, Log−normal probability plot

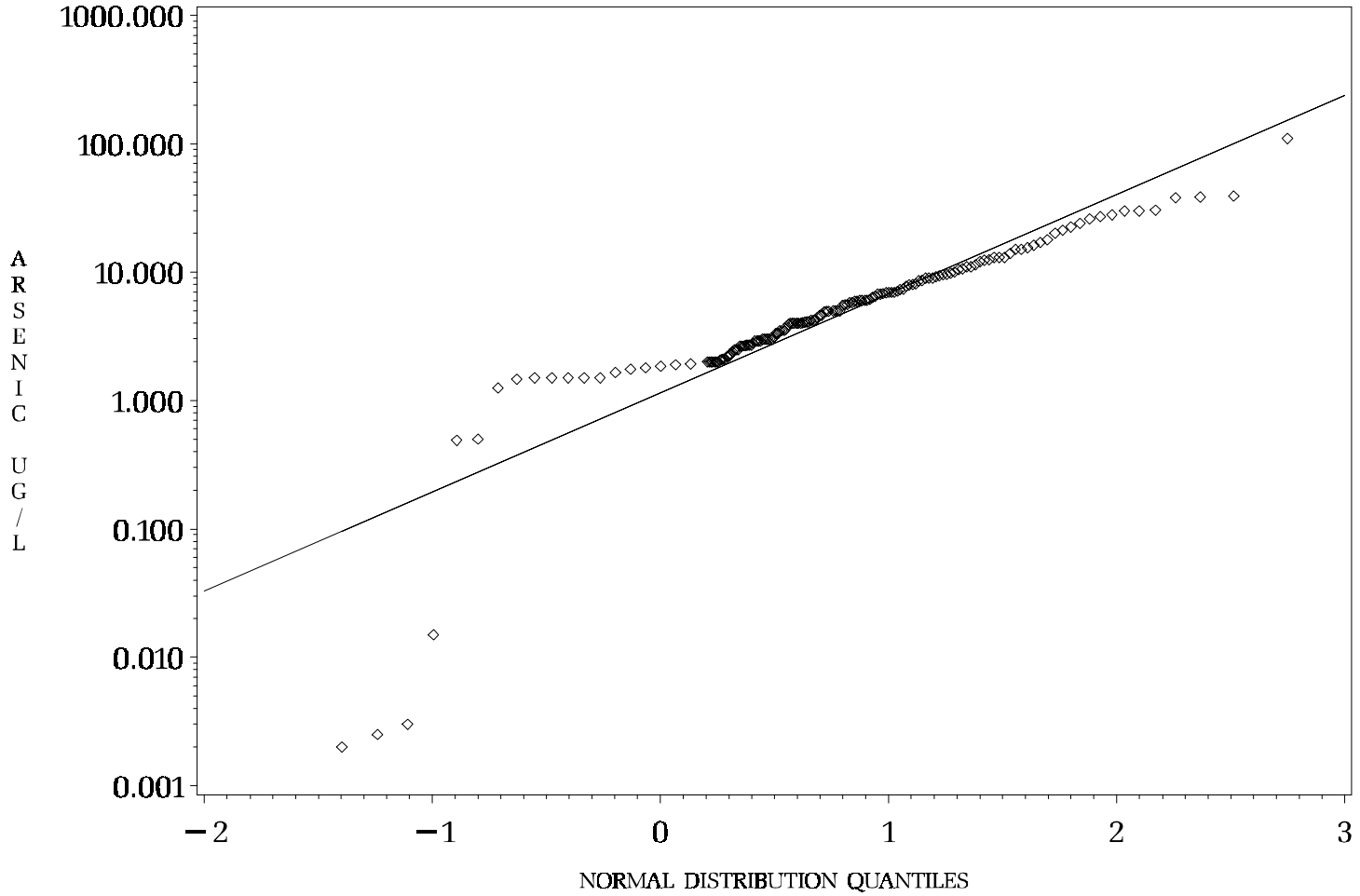


Figure B−51: System means of NTNCWS GW arsenic concentrations for IN, Log−normal probability plot

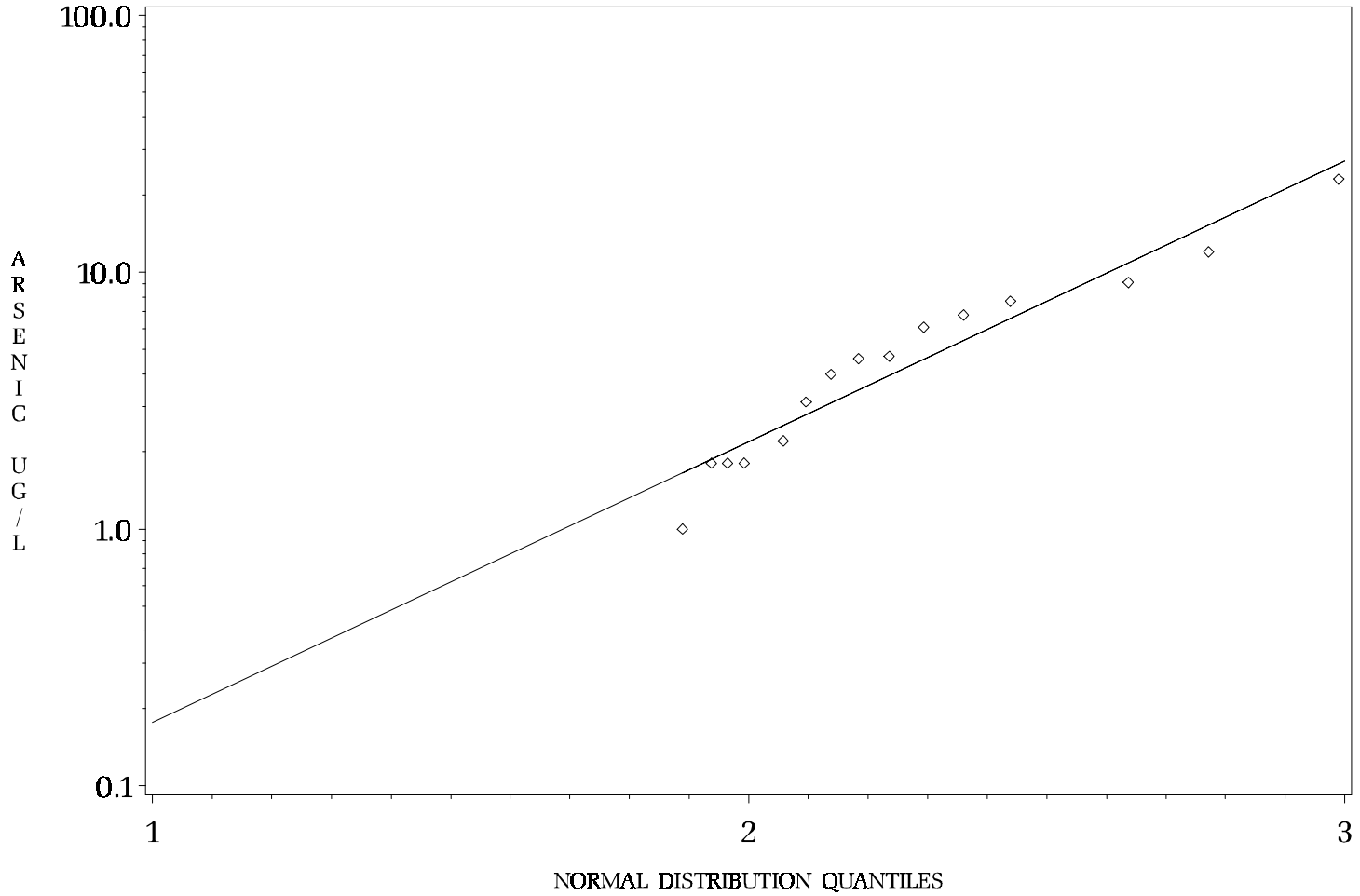


Figure B−52: System means of NTNCWS GW arsenic concentrations for KS, Log−normal probability plot

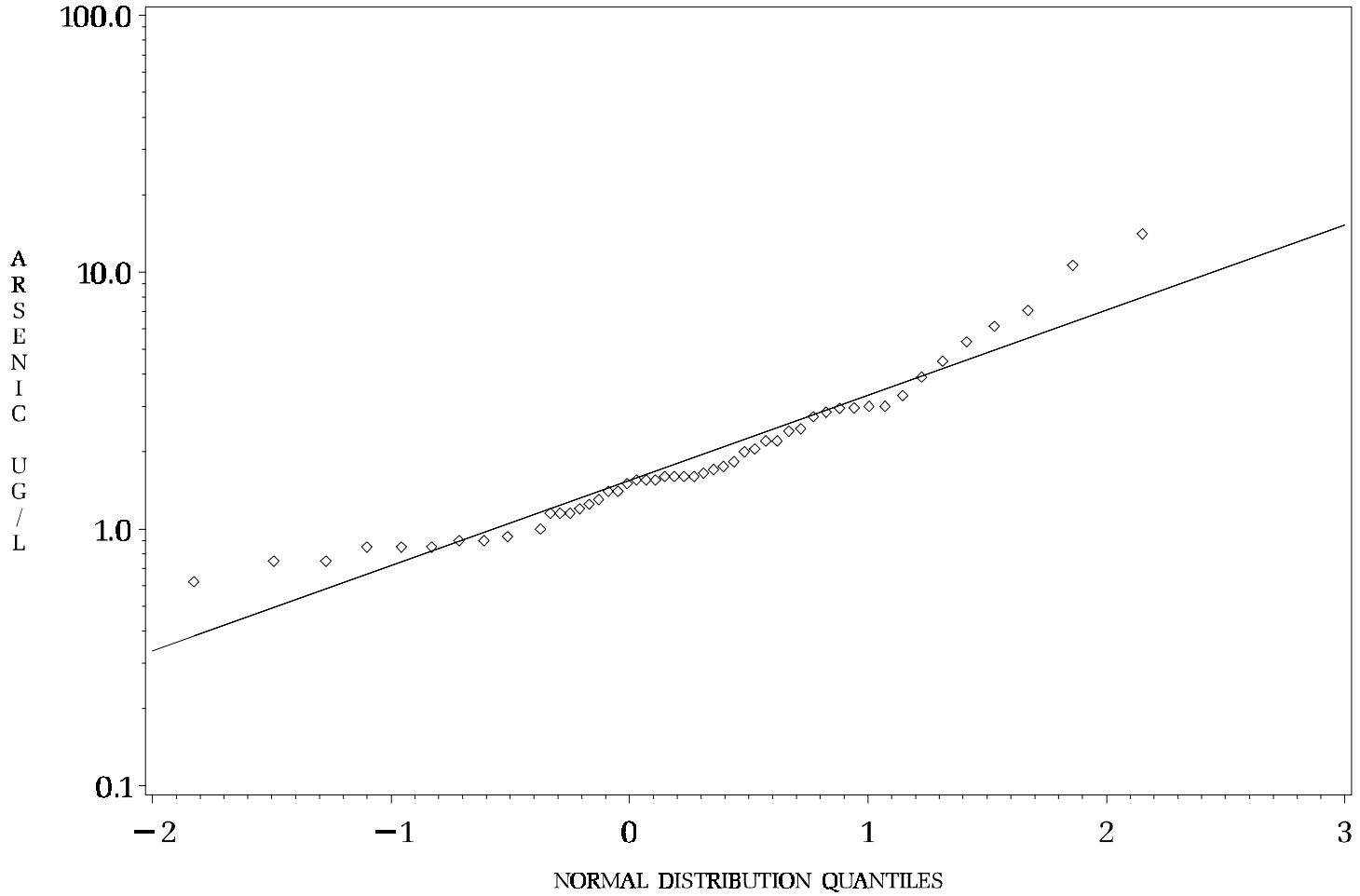


Figure B−53: System means of NTNCWS GW arsenic concentrations for MI, Log−normal probability plot

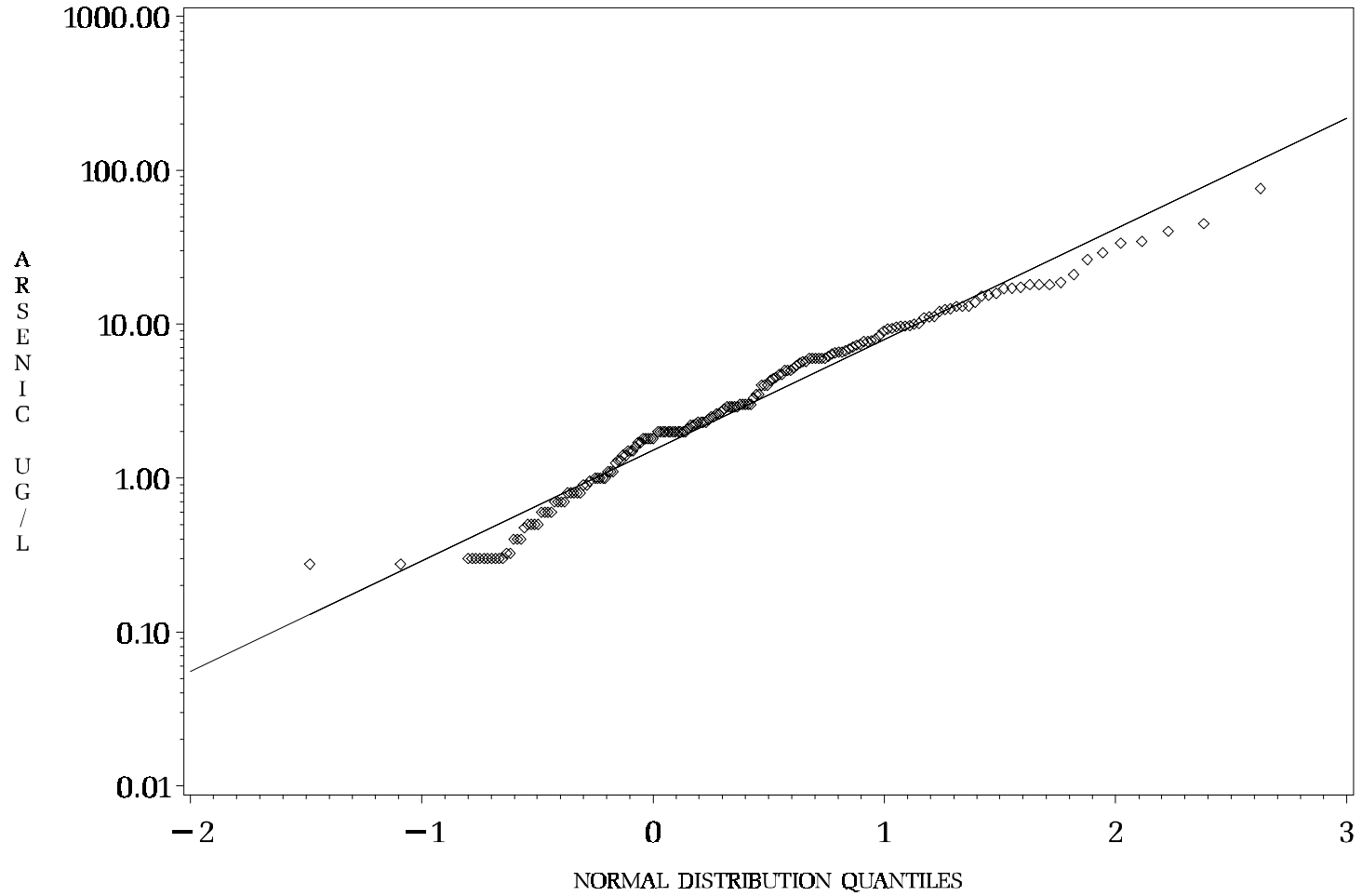


Figure B−54: System means of NTNCWS GW arsenic concentrations for MN, Log−normal probability plot

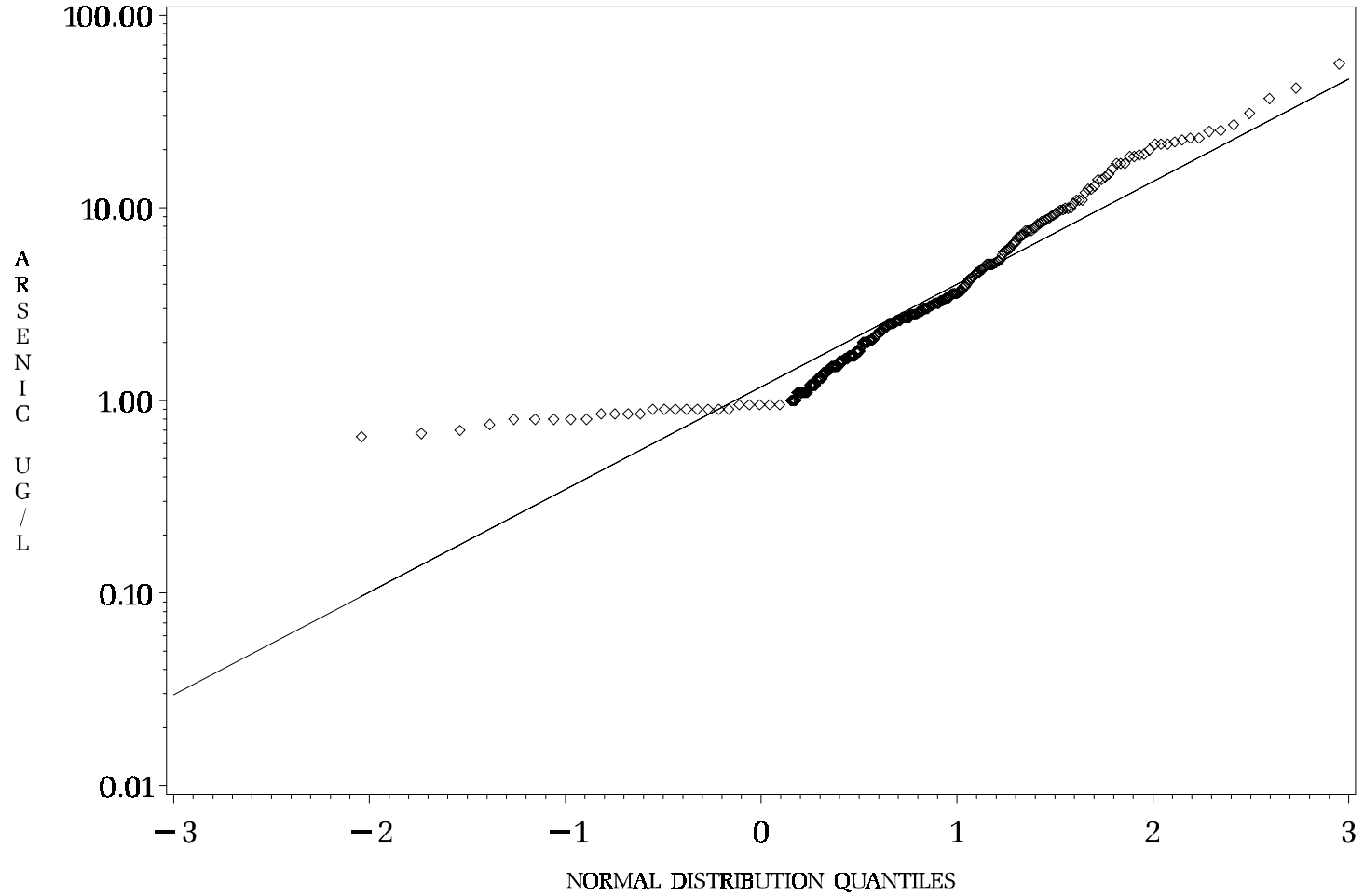


Figure B−55: System means of NTNCWS GW arsenic concentrations for NC, Log−normal probability plot

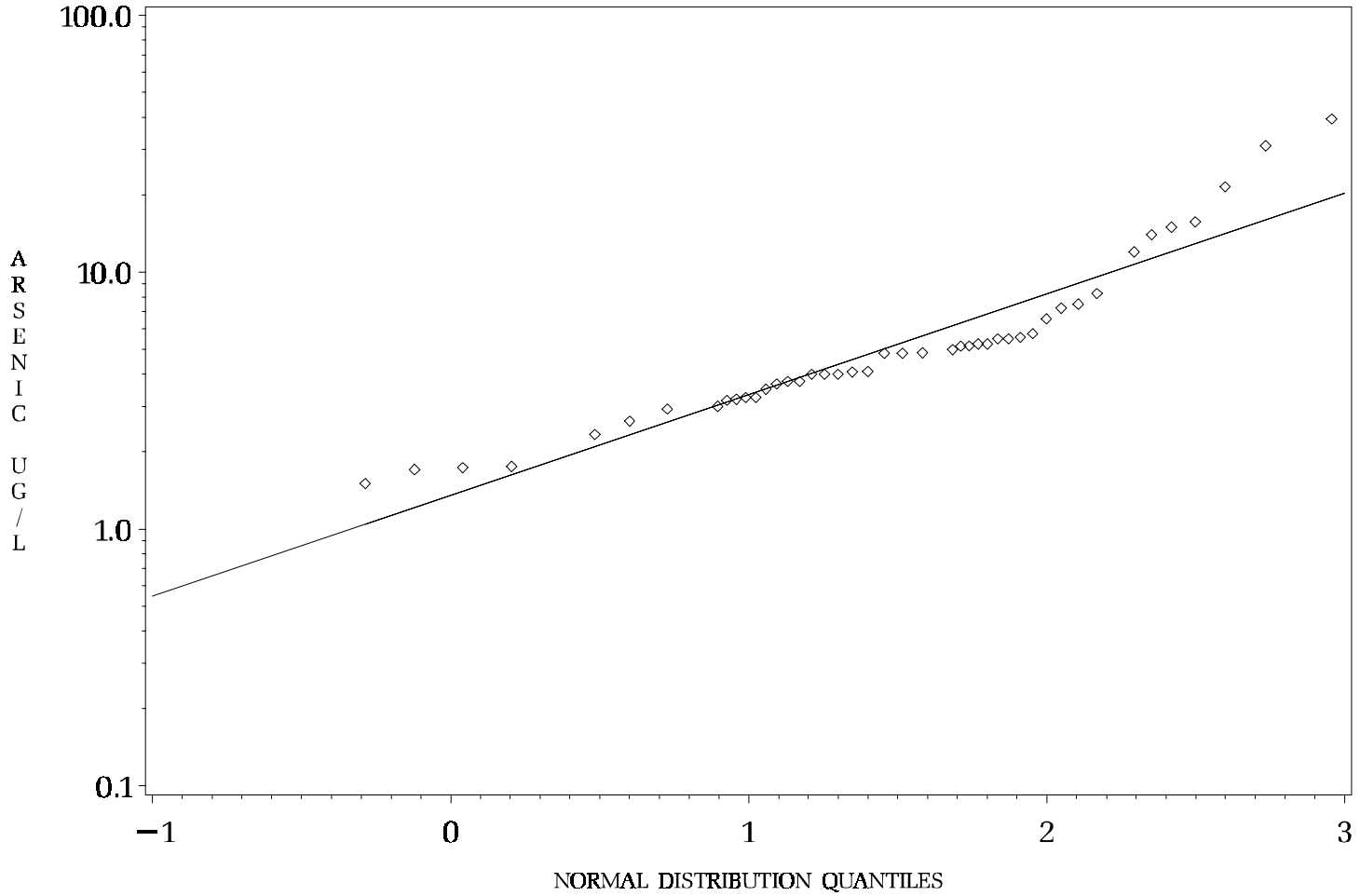


Figure B−56: System means of NTNCWS GW arsenic concentrations for ND, Log−normal probability plot

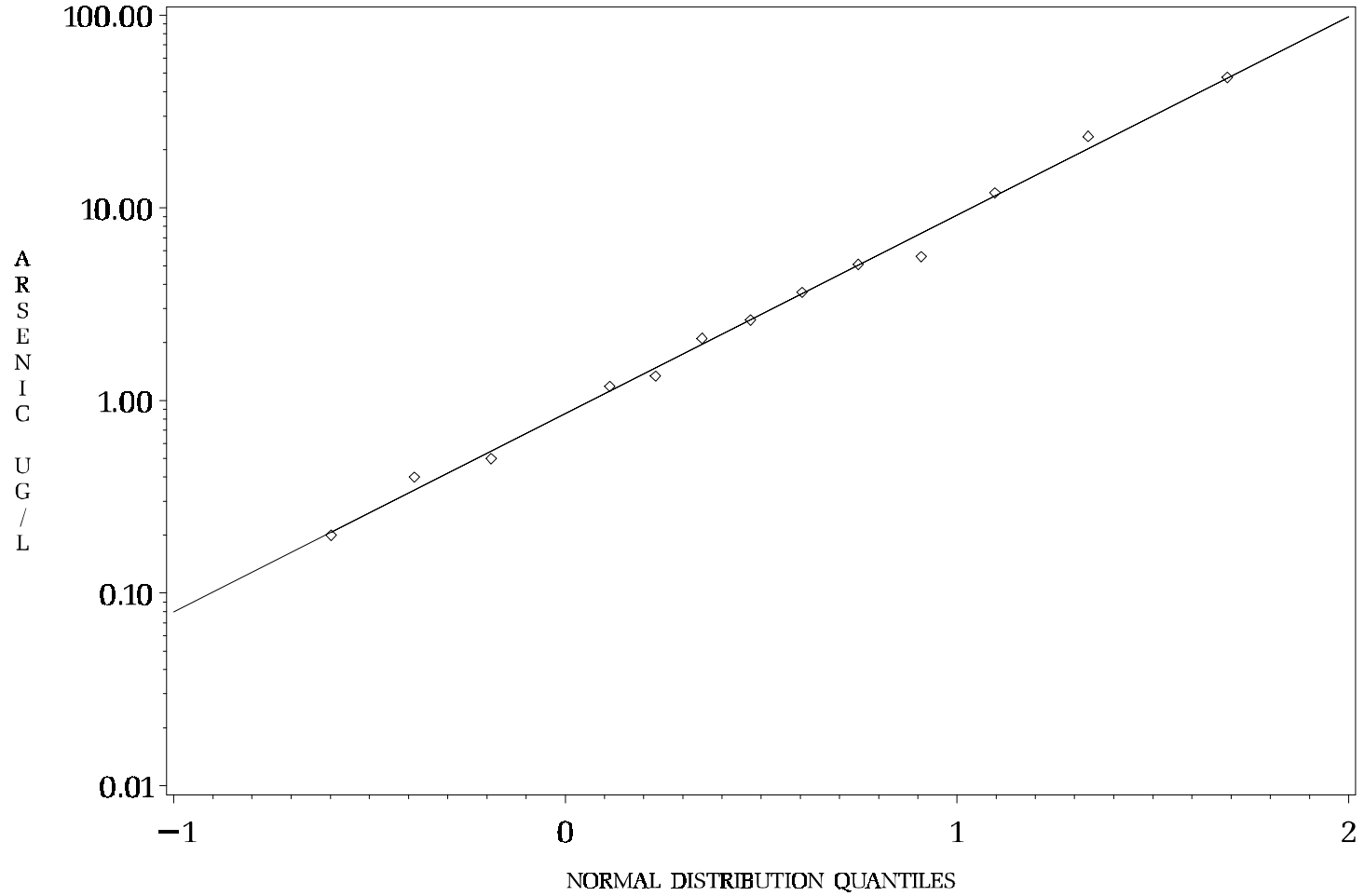


Figure B−57: System means of NTNCWS GW arsenic concentrations for NJ, Log−normal probability plot

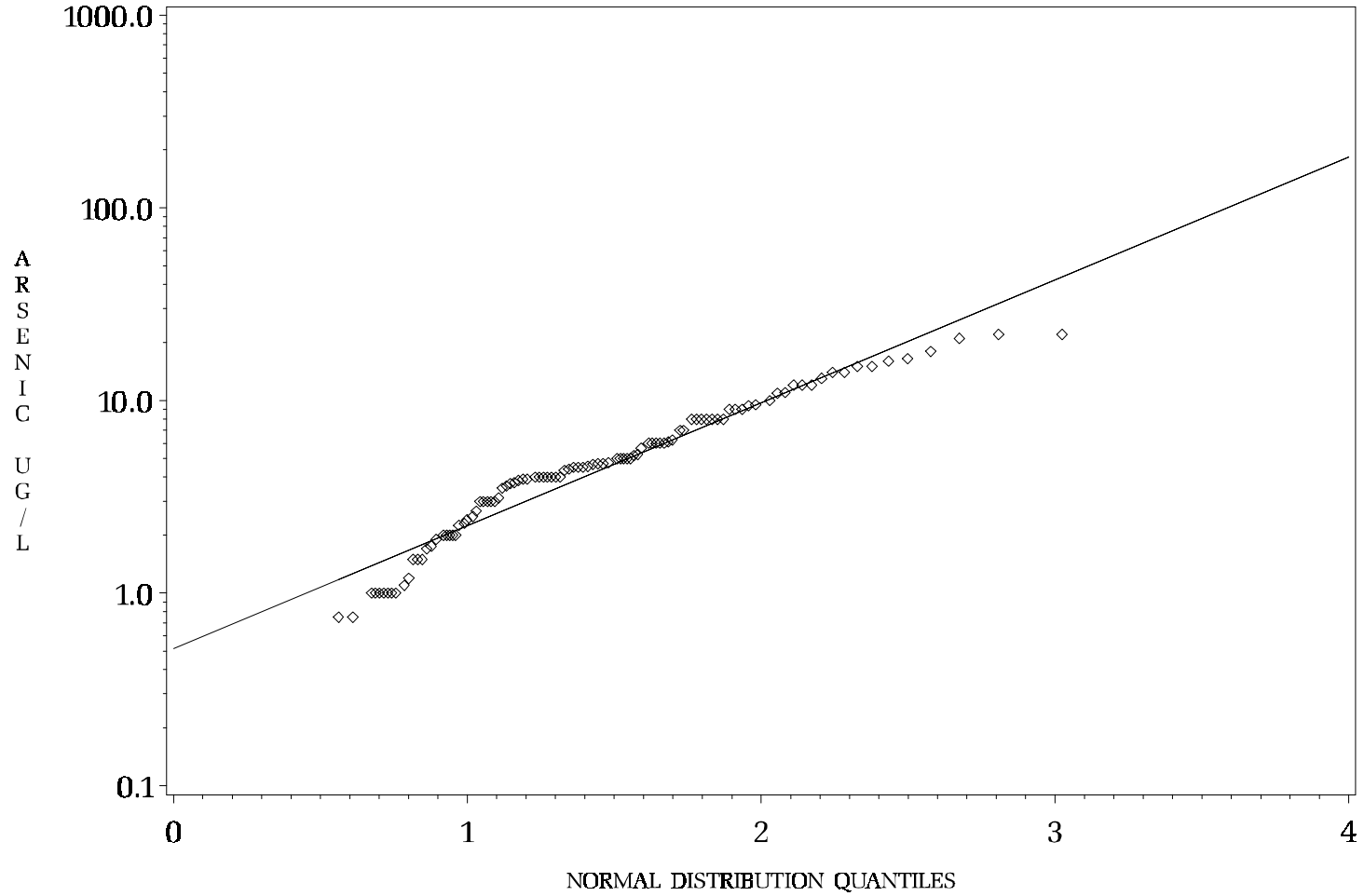


Figure B−58: System means of NTNCWS GW arsenic concentrations for NM, Log−normal probability plot

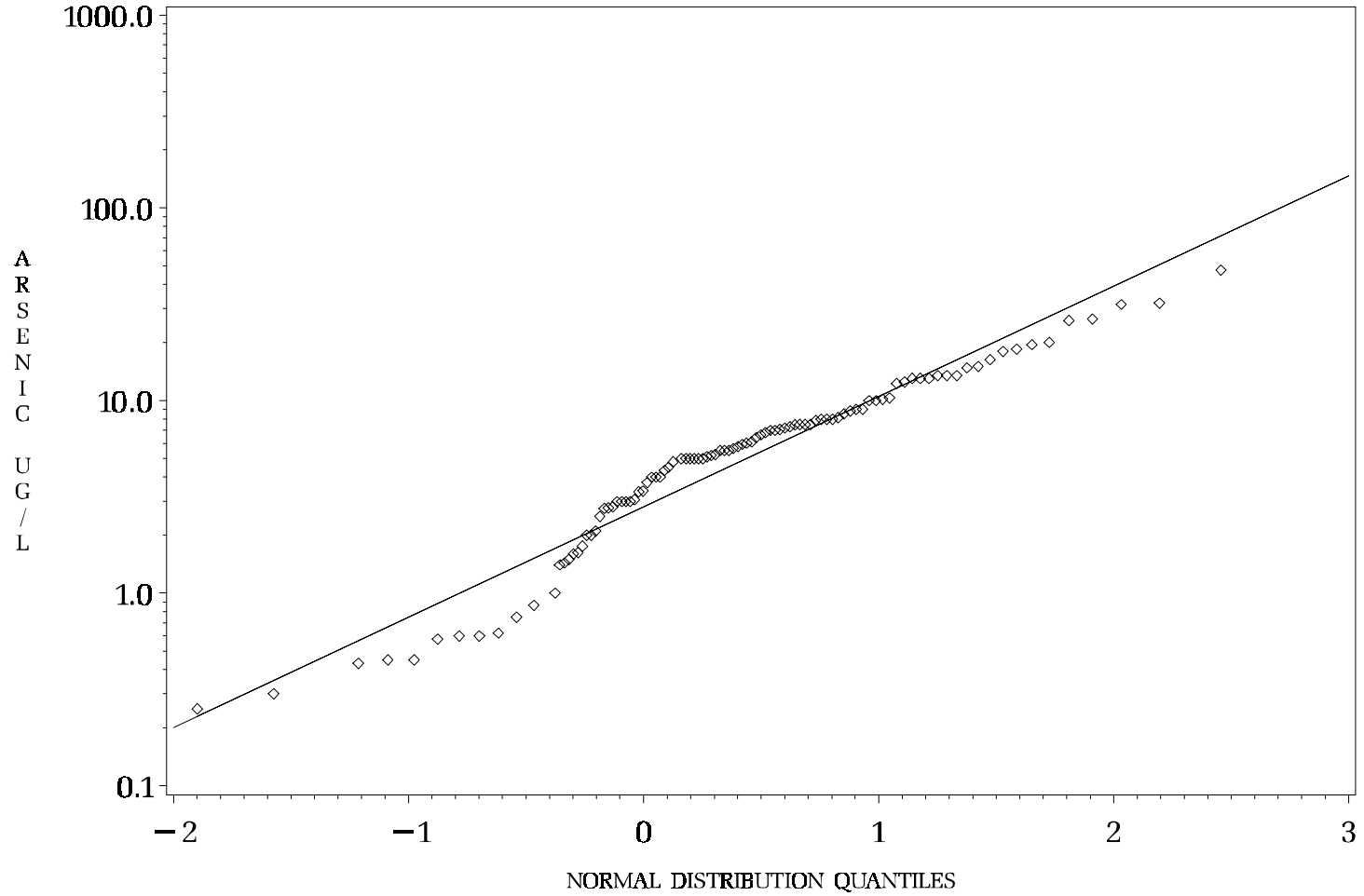


Figure B−59: System means of NTNCWS GW arsenic concentrations for OR, Log−normal probability plot

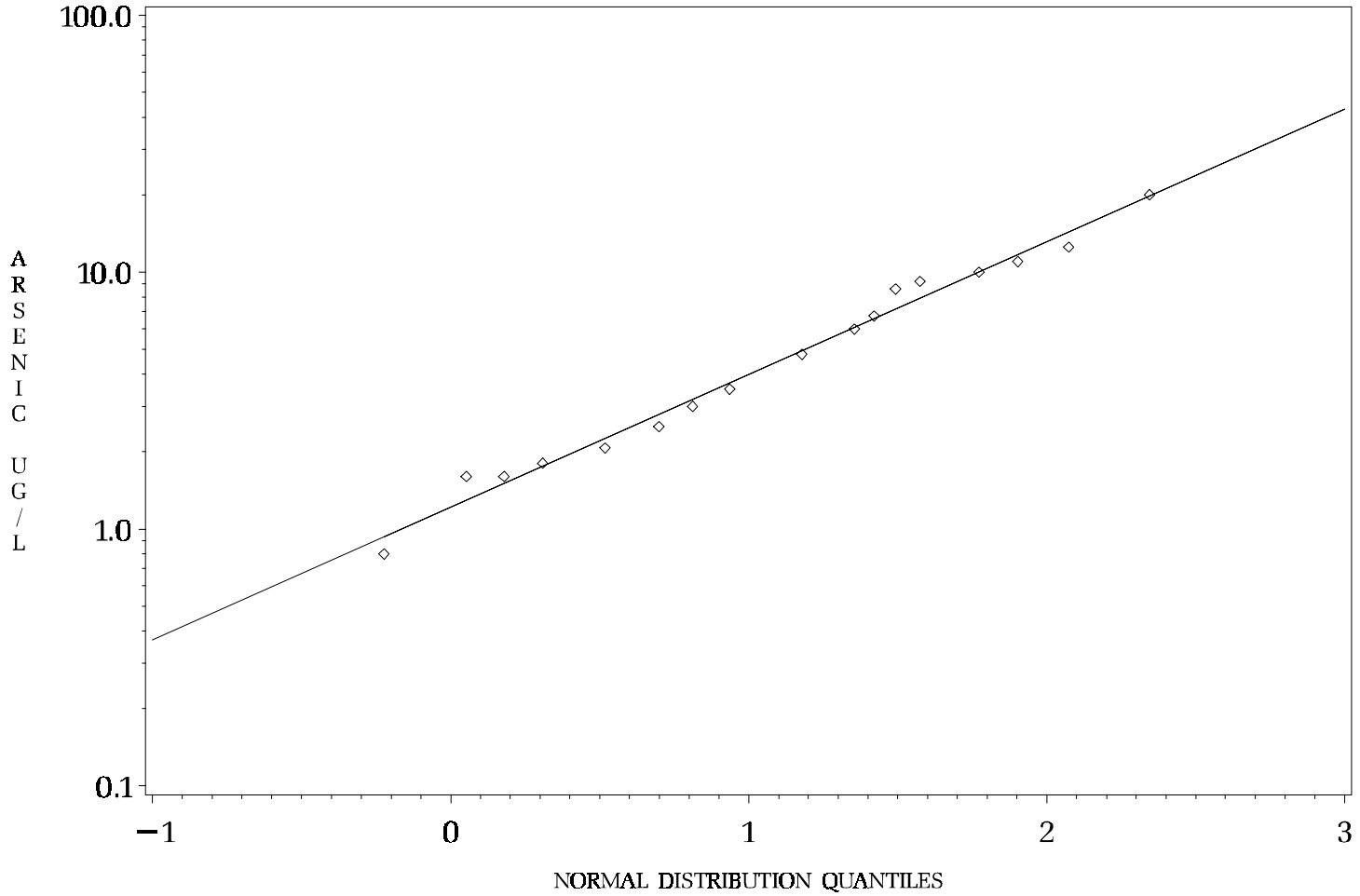


Figure B−60: System means of NTNCWS GW arsenic concentrations for TX, Log−normal probability plot

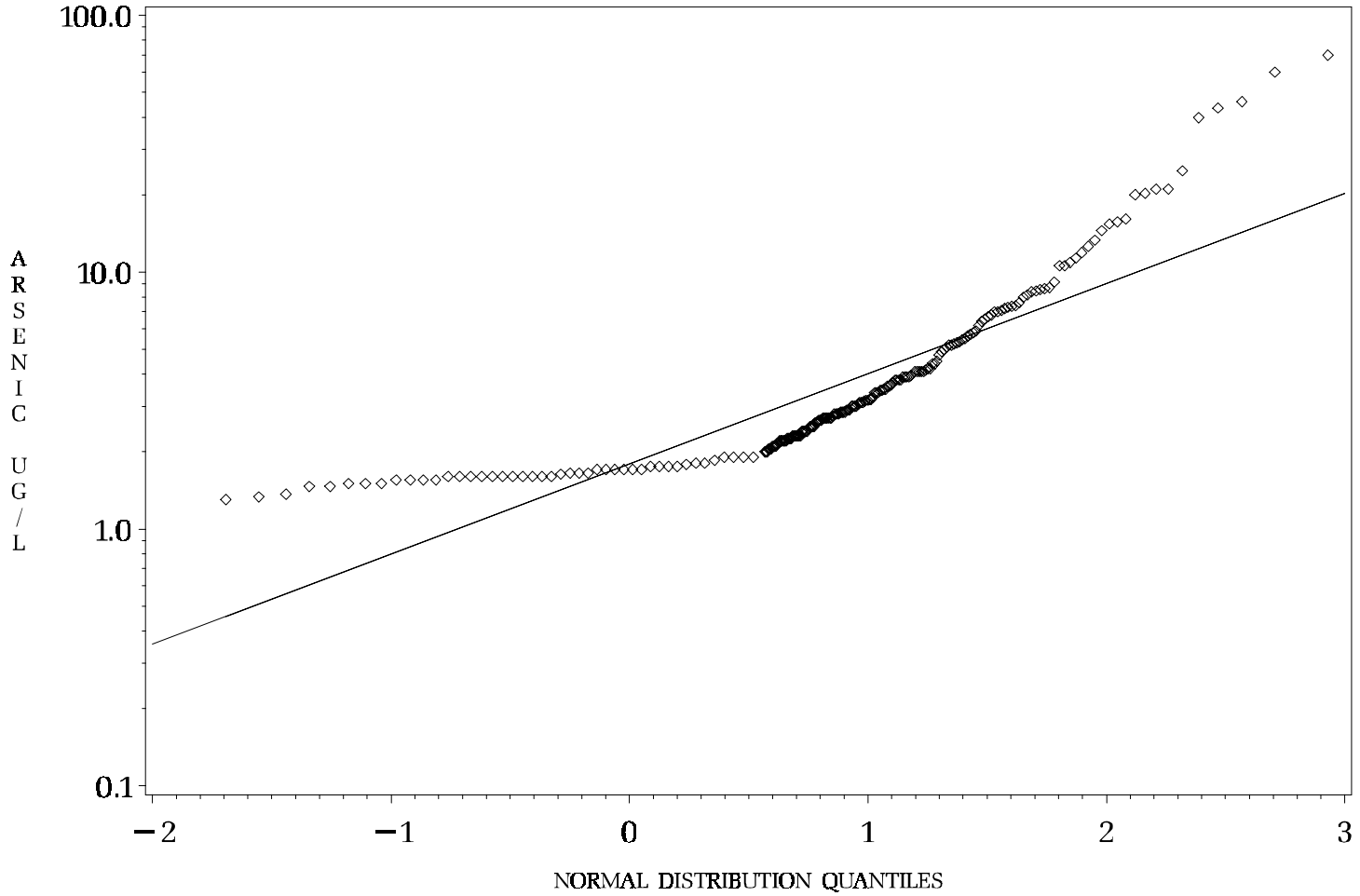


Figure B−61: System means of NTNCWS GW arsenic concentrations for UT, Log−normal probability plot

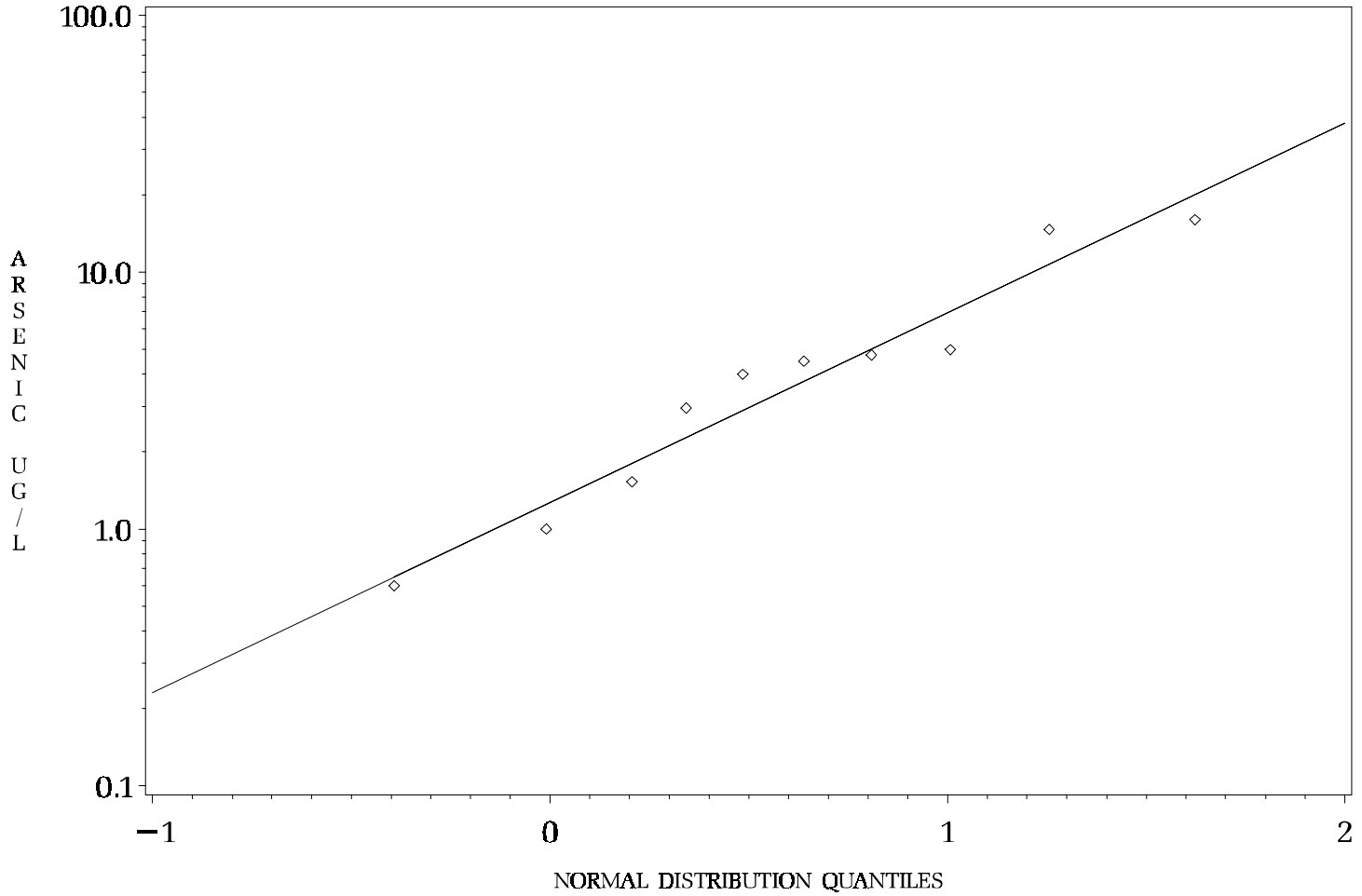


Figure B−62: System means of NTNCWS SW arsenic concentrations for AK, Log−normal probability plot

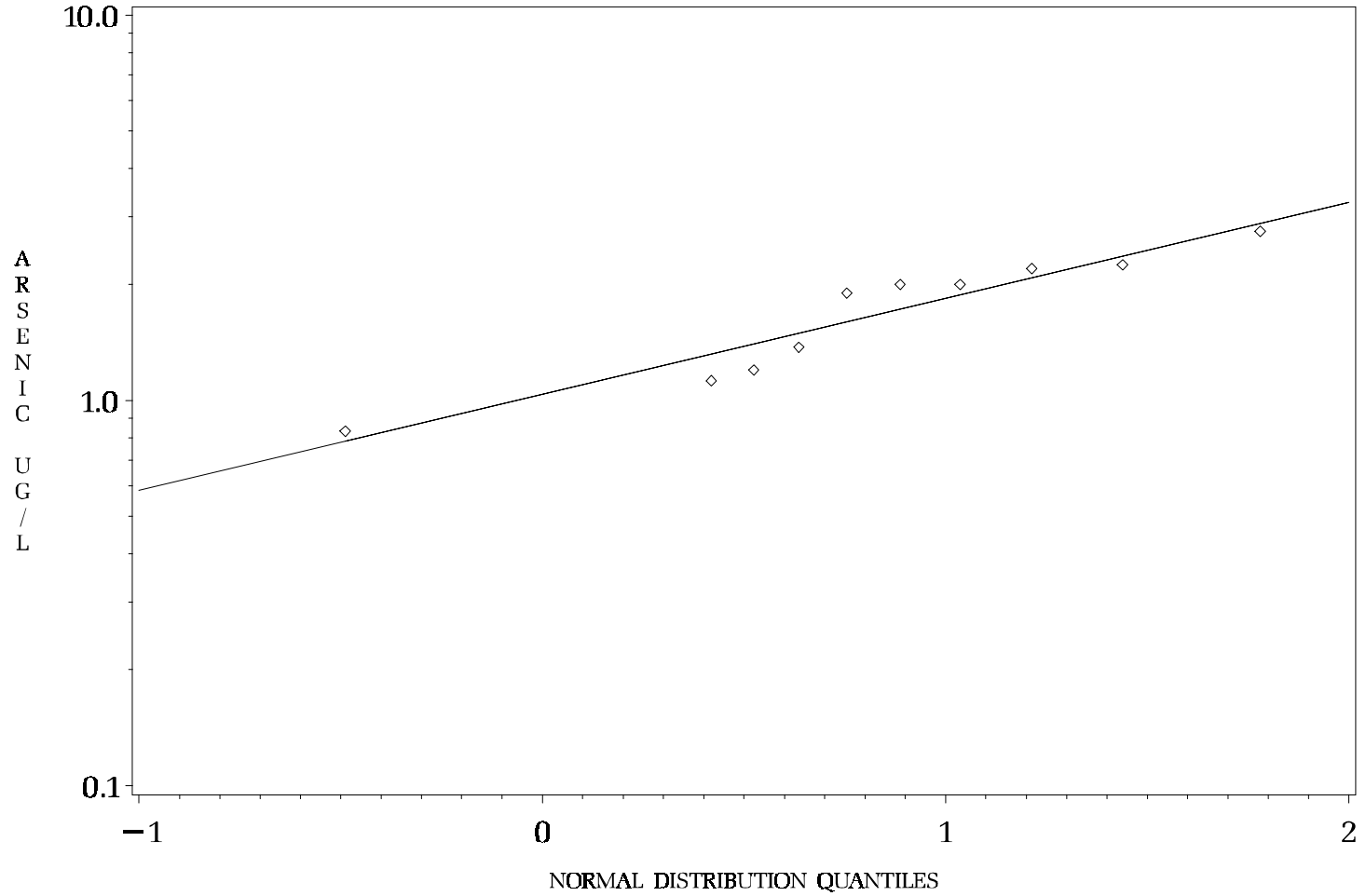


Figure B−63: System means of NTNCWS SW arsenic concentrations for ND, Log−normal probability plot

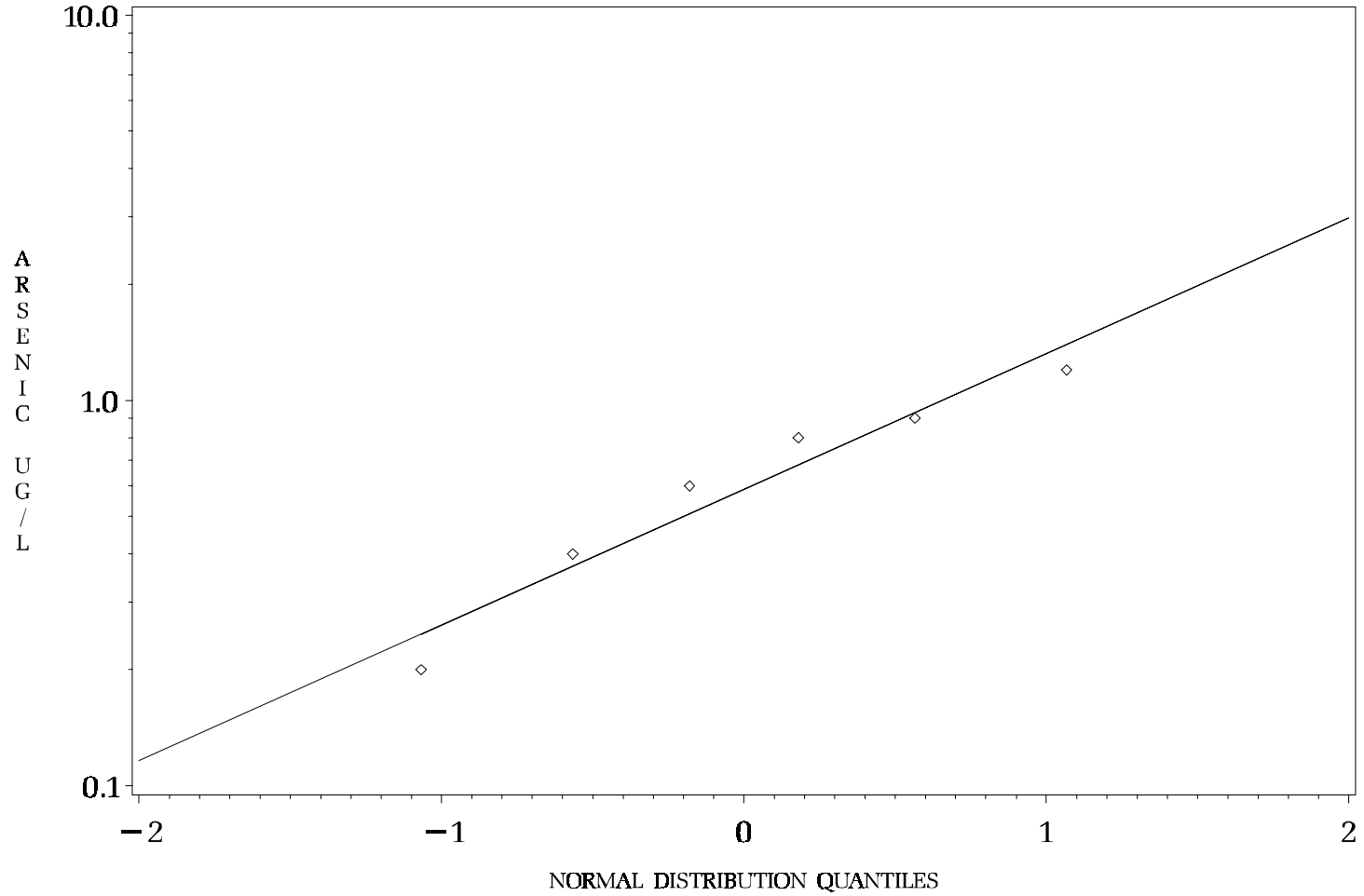
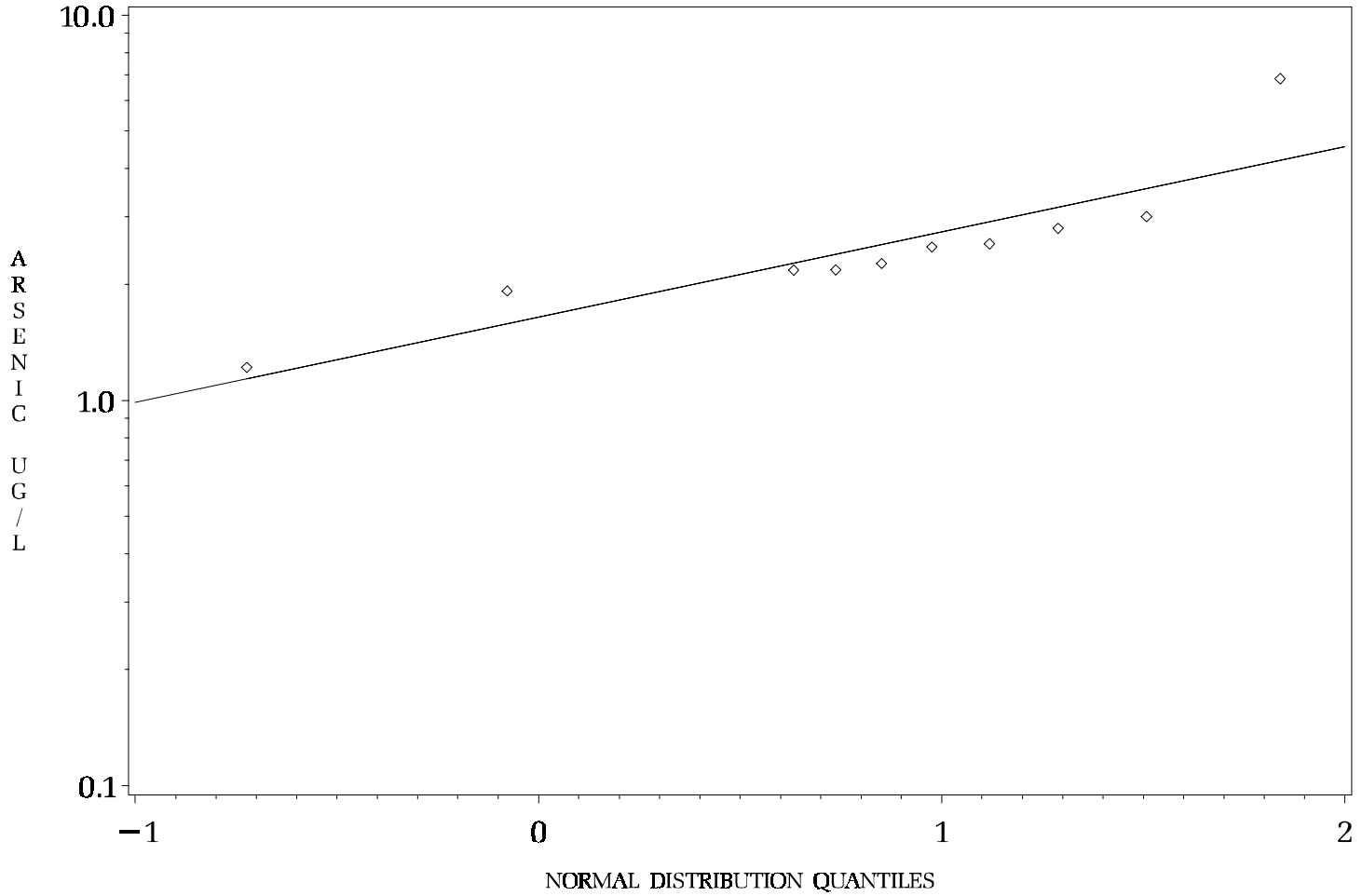


Figure B−64: System means of NTNCWS SW arsenic concentrations for TX, Log−normal probability plot



Appendix C
Summaries of Pre-1980 Data Sets

This page intentionally left blank

Arsenic Occurrence in Public Water Supplies
Reported by the 1969 Community Water Supply Survey

Ground Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	366	20	95%	15.0	64.0
501-3,301	148	2	99%	30.0	100.0
3,301-10,000	81	7	91%	7.5	30.0
10,001-100,000	69	2	97%	15.0	30.0
> 100,000	9	2	78%	5.0	10.0
Total	673	33	95%	5.0	100.0

Surface Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	31	1	97%	30.0	30.0
501-3,301	38	4	89%	30.0	30.0
3,301-10,000	15	1	93%	30.0	30.0
10,001-100,000	14	3	79%	15.0	30.0
> 100,000	8	0	0%	NA	NA
Total	106	9	92%	15.0	30.0

NA: Not applicable, no positive detections were reported.

Arsenic Occurrence in Public Water Supplies
Reported by the 1978 Community Water Supply Survey

Ground Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	120	22	82%	2.5	28.0
501-3,301	85	16	81%	3.2	17.0
3,301-10,000	34	5	15%	3.5	17.3
10,001-100,000	20	6	70%	3.1	8.2
> 100,000	0	NA	NA	NA	NA
Total	259	49	82%	2.5	28.0

Surface Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	17	0	100%	NA	NA
501-3,301	36	1	97%	2.5	2.5
3,301-10,000	21	0	100%	NA	NA
10,001-100,000	20	2	90%	4.4	10.7
> 100,000	0	NA	NA	NA	NA
Total	94	3	92%	2.5	10.7

NA: Not applicable.

Arsenic Occurrence in Public Water Supplies
Reported by the Rural Water Survey

Ground Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	18	8	56%	5.0	82.0
501-3,301	38	12	68%	2.0	40.0
3,301-10,000	8	2	25%	3.0	6.0
10,001-100,000	5	1	80%	8.0	8.0
> 100,000	2	0	0%	NA	NA
Total	71	23	68%	2.0	82.0

Surface Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	0	0	NA	NA	NA
501-3,301	3	0	100%	NA	NA
3,301-10,000	7	0	100%	NA	NA
10,001-100,000	4	1	75%	3.0	3.0
> 100,000	7	1	86%	5.0	5.0
Total	21	2	92%	3.0	5.0

NA: Not applicable, no positive detections were reported.

Arsenic Occurrence in Public Water Supplies
Reported by the National Organics Monitoring Survey

Ground Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	0	NA	NA	NA	NA
501-3,301	0	NA	NA	NA	NA
3,301-10,000	0	NA	NA	NA	NA
10,001-100,000	3	2	67%	7.0	10.0
> 100,000	12	4	33%	5.0	18.0
Total	15	6	60%	5.0	18.0

Surface Water Supplies					
Population Served	**Number of Samples**	**Number of Detects**	**Percent Nondetects**	**Minimum (µg/L)**	**Maximum (µg/L)**
25-500	1	0	100%	NA	NA
501-3,301	0	NA	NA	NA	NA
3,301-10,000	3	0	100%	NA	NA
10,001-100,000	17	6	65%	5.0	20.0
> 100,000	65	13	80%	5.0	17.0
Total	86	19	78%	5.0	20.0

NA: Not applicable, no positive detections were reported.

Appendix D
Database Specifications and Data Conditioning

This page intentionally left blank

Appendix D-1
Individual State Database Specifications
for Preliminary Database
(see Section 4.1.3 for further modifications)

This page intentionally left blank

ARSENIC DATA CONDITIONING NOTES FOR OCCURRENCE AND EXPOSURE DATABASES (AOED, GRAND.CPT, and INTRA.CPT)

DATA CONVENTIONS:
All individual records are identified by the sampling point id number *(e.g.*, S3519) or result id *(e.g.*, R3519) these numbers were uniquely assigned to every record received for tracking purposes.

LIST OF DATABASES:
NIRS, NAOS, Alabama, Alaska, Arizona, Arkansas, California, Illinois, Indiana, Kansas, Kentucky, Maine, Michigan, Minnesota, Missouri, Montana, Nevada, New Hampshire, New Jersey, New Mexico, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Utah, and Texas.

NIRS (National Inorganics and Radionuclides Survey)

Data received from EPA.

State:	Data given (12/18/97).
County:	No data given (12/18/97).
PWSID:	Data obtained and transferred from the NIRS2 database. Two missing PWSIDs were obtained from SDWIS. Some PWSIDs beginning with '04' are tribal systems. Set Ocala PWSID to NIRS0 (4/2/98).
Type of WSS:	No data given (12/18/97). Data obtained from documentation *(Arsenic Occurrence: USEPA Seeks Clearer Picture*, 1994) provided by the EPA WAM (2/4/98).
Source Type:	No data given. (12/18/97) Data obtained from documentation *(Arsenic Occurrence: USEPA Seeks Clearer Picture*, 1994) provided by the EPA WAM (2/4/98).
PWS Name:	Used city name (12/18/97) in this field is PWS name found in SDWIS (3/4/98).
Population:	Population numbers for four quarters and an annual average were given. The average was used (12/18/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	Obtained from documentation *(Occurrence Assessment for Arsenic in Public Drinking Water Supplie*s, September, 1992).
Sample Type:	Assumed (2/4/98).
Sample Date:	Data given (12/18/97).
EPA Analytical Value:	Data given (12/18/97).
Reporting Limit:	Obtained from documentation *(Occurrence Assessment for Arsenic in Public Drinking Water Supplie*s, September, 1992) provided by the EPA WAM.

Conditioning Notes:

- Tribal systems are included in this data set.
- Date Range: 9/84–10/86.
- Used NIRS data set combined with NIRS2 PWSIDs to obtain missing data elements.

NAOS (National Arsenic Occurrence Survey)

Data received from EPA

State:	Data given (12/18/97).
County:	No data given (12/18/97).
PWSID:	No PWSIDs were given (12/18/97). Added "NAOS"& sample id (4/3/98).
Type of WSS:	No data given (12/18/97).
Source Type:	Data converted from a numeric format to proper database code.
PWS Name:	No data given (12/18/97).
Population:	Data given (12/18/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	Obtained from documentation (*National Compliance Assessment and Costs for the Regulation of Arsenic in Drinking Wate*r, January, 1997) provided by the EPA WAM (2/4/98).
Sample Type:	Data given (12/18/97).

Sample Date: No data given (12/18/97).
EPA Analytical Value: Data given (12/18/97).
Reporting Limit: Data given (12/18/97).
Detection: Assigned based on reporting limit. Values greater than the reporting limit qualified as Detects.

Conditioning Note:

- Large CWS systems represented.

Alabama

Data received from Ed Thomas at EPA, originally from Tom DeLoach, AL

State: ISSI generated for database purposes.
County: Provided in database.
PWSID: Provided in database.
Type of WSS: Data obtained from SDWIS.
Source Type: Data obtained from SDWIS.
PWS Name: Provided in database.
Population: Data obtained from SDWIS.
Sampling Point ID: Provided in database.
Sampling Point Type: Provided in database.
Sample Type: Provided in database.
Sample Date: Provided in database.
EPA Analytical Value: Provided in database.
Reporting Limit: Non-detects indicated with "0" result. Conducted a frequency analysis of detects, and found that the lowest commonly occurring value was 1 ppb. Assumed RL = 1 ppb.
Detection: Provided in database.

Alaska

Data received from SRA

State: Added "AK" to all records (12/20/97).
County: No data given (12/20/97).
PWSID: Data given but added "AK2" to given sysid (12/20/97).
Type of WSS: Data obtained from SDWIS (12/20/97).
Source Type: Data given but had to be modified in the following way: **S, A,Y,** P="SW" **C, W,** G="GW" (12/20/97).
PWS Name: Data given (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98).
Sampling Point Type: Data provided by the data contact (12/20/97).
Sample Type: Assumed (12/20/97).
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data provided by the data contact.
Detection: Data given (12/20/97).

Conditioning Notes:

- Date range: 12/66–4/97.
- Deleted records S37091 and S37225 (detects reported at zero)

Arizona

Data received from Linda Bragg, Az. Dept. of Env. Quality

State:	Added "AZ" to all records.
County:	Data given.
PWSID:	Added "AZ04" before the given system ID.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Data given.
Population:	Data obtained from SDWIS.
Sampling Point ID:	ISSI generated for database purposes (5/27/98)
Sampling Point Type:	No data given.
Sample Type:	No data given.
Sample Date:	Data given.
EPA Analytical Value:	Data given.
Reporting Limit:	Data provided by data contact. Data prior to 1988 were sent with no indication of positive or negative detection. The data (2,450 records) were moved to the Obsolete data file. The data provided had "=" sign, which mean the values, were confirmed as positive results. For PWS AZ0413154 (10/26/97), the arsenic value reported was >25 ppb.
Detection:	No data given.

Conditioning Notes:

- Data were given in four separate files--PWS information before 1/1/93; PWS information after 1/1/93; results before 1/1/93; and result after 1/1/93. These files were conditioned to create properly formatted spreadsheets. Several clues indicated that dates were only given for the first sample taken on each day. In addition, PWSIDs and PWS Names were only given for the first sample for each PWS. To create a properly formatted spreadsheet, the dates, PWSIDs and PWS Names were copied to their correct results/records.
- Date range: 01/88-4/98
- Deleted S150047, S150049, and S151149 (zero results reported for detections). (FM 8/3/98)

Arkansas

Data received from Tom Poeten, EPA Region 6

State:	ISSI generated for database purposes.
County:	Provided in database.
PWSID:	Database provided abbreviated PWSID numbers, which were converted to complete PWSID numbers in accordance with directions from Tom Poeten.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Provided in database.
Population:	Data obtained from SDWIS.
Sampling Point ID:	Provided in database.
Sampling Point Type:	Provided in database.
Sample Type:	Provided in database.
Sample Date:	Provided in database.
EPA Analytical Value:	Provided in database.
Reporting Limit:	Provided in database.
Detection:	Provided in database.

California

Data received from SRA and verified with CA contacts.

State:	Added "CA" to all records.
County:	Data given (12/20/97).
PWSID:	Data given (12/20/97).
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Data obtained from SDWIS.
Population:	Data obtained from SDWIS.
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	No data given (12/20/97).
Sample Type:	Data given (12/20/97).
Sample Date:	Data given (12/20/97).
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data obtained from the data contact. (3/5/98)
Detection:	Data given as "<" in the "XMOD" column (3/5/98)

Conditioning Notes:

- Date range: 11/1/90 –5/22/97.
- Records prior to November 1, 1990 were deleted (5/29/98). The deleted records are located in the Obsolete file.
- Originally, the raw data from californ.dbf was used. After a close analysis and conversations with California representatives, it was determined that arsenic.dbf was more representative of the California arsenic data, and the California data were re-analyzed.
- Reporting Limit Information was not available for records prior to 1990, so 12687 records were removed from IAOED, representing 1120 PWS and all systems sizes. 19 PWS with less than 25 people served, 57 PWS with 25-100 people served, 121 PWS with 100-500 people served, 124 PWS with 500-1000 people served, 253 PWS with 1000-3300 people served, 209 PWS with 3300-10000 people served, 228 PWS with 10000-50000 people served, 58 PWS with 50000-100000 people served, 49 PWS with 100000-1 million people served, and 2 PWS with greater than 1 million people served.
- Records labeled unknown were deleted from the database.

Illinois

Data received from Ed Thomas EPA, Mike Crumly, IL

State:	ISSI generated for database purposes.
County:	Provided in database.
PWSID:	Provided in database.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Provided in database.
Population:	Data obtained from SDWIS.
Sampling Point ID:	Provided in database.
Sampling Point Type:	Provided in database.
Sample Type:	Provided in database.
Sample Date:	Provided in database.
EPA Analytical Value:	Provided in database.
Reporting Limit:	Provided in database.
Detection:	Provided in database.

Indiana

Data received from Al Lao, Phil Zellinger, State of Indiana

State:	ISSI generated for database purposes.
County:	Provided in database.
PWSID:	Provided in database.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Provided in database.
Population:	Data obtained from SDWIS.
Sampling Point ID:	Provided in database.
Sampling Point Type:	Provided in database.
Sample Type:	Provided in database.
Sample Date:	Provided in database.
EPA Analytical Value:	Provided in database.
Reporting Limit:	Method detection numbers provided for later results. Phil Zellinger (IN) provided associated detection limits for each method number. Earlier results (pre 1996) reported as positive at the reporting level. For these early samples, could not discriminate between detects and non-detects, so samples collected before 1996 were omitted from the database.
Detection:	Provided in database.

Kansas

Data received from Bob Bostrom, Kansas Dept. of Health and Environment

State:	Added "KS" to all records (12/20/97).
County:	Data given (12/20/97).
PWSID:	Data given (12/20/97).
Type of WSS:	Data given but had to be modified in the following way: **NTN, TNC** = "NTNCWS", and **CAP, CFF, CIN, CMH, CMU, CPM, CPV, CRW, CSC, CSI, CWD, CWS, CWW, MHP** = "CWS" (12/20/97)
Source Type:	Data given but had to be modified in the following way: **G, W** = "GW" **S, P** = "SW" (12/20/97).
Population:	Data given (12/20/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98)
Sampling Point Type:	Data provided by the data contact.
Sample Type:	Data given (12/20/97).
Sample Date:	Data given (12/20/97).
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data provided by data contact.
Detection:	Assigned based on reporting limit.

Conditioning Notes:

- Records with sampling point IDs S22843, S23461, S25050, and S25996 were deleted because they had no PWSIDs.
- Date range: 1/91–12/97

Kentucky

Data received from EPA in the earlier "11 States" data set

State:	Data given (12/21/97).
County:	No data given (12/21/97).
PWSID:	Data given in 5-, 6-, 7- digit format. (12/21/97) used SDWIS to obtain correct PWSIDs. Some PWSID were still invalid, the data from MI, CA, ID are invalid (3/19/98).
Type of WSS:	No data given from raw table (12/21/97). Stakeholders information provided this data (all data=CWS) [4/2/98]. Linked with SDWIS to determine TWSS for active systems.
Source Type:	Provided converted values to IAOED standards (12/21/97).
PWS Name:	Used city name, does not correlate with SDWIS system name
Population:	Data given (12/21/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	EPA WAM provided data (4/2/98).
Sample Type:	Assumed (4/2/98).
Sample Date:	No Data given.
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data obtained from state contacts. (3/20/98)
Detection:	Data given (12/20/97).

Conditioning Notes:

- No dates provided for Kentucky, however, data was collected in 1993–1994.
- The data contained in this data set was already modified and combined by EPA.

Maine

Data received from EPA in the earlier "11 States" data set

State:	Data given (12/21/97).
County:	No data given (12/21/97).
PWSID:	Data given in 5-, 6-, 7- digit format. (12/21/97) used SDWIS to obtain correct PWSIDs. Some PWSID were still invalid, the data from MI, CA, ID are invalid (3/19/98).
Type of WSS:	No data given from raw table (12/21/97). Stakeholders information provided this data (all data=CWS) [4/2/98]. Linked with SDWIS to determine TWSS for active systems.
Source Type:	Provided converted values to IAOED standards (12/21/97).
PWS Name:	Used city name, does not correlate with SDWIS system name
Population:	Data given (12/21/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	EPA WAM provided data (4/2/98).
Sample Type:	Assumed (4/2/98).
Sample Date:	No Data given.
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data obtained from state contacts. (3/20/98)
Detection:	Data given (12/20/97).

Conditioning Note:

- The data contained in this data set was already modified and combined by EPA.

Michigan

Data received from Mark Breithart, MI DEQ.

State:	Added "MI" to all records (02/08/98).
County:	Data given (3/3/98).
PWSID:	Data given (12/20/97).
Type of WSS:	Data given (12/20/97).
Source Type:	Data given (12/20/97).

PWS Name: Data given (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98)..***Sampling Point Type:*** Data given (for details, see Michigan documentation).
Sample Type: Data given (12/20/97).
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data given (12/20/97).
Detection: Data given (12/20//97).

Conditioning Notes:
- Date range: 1/83–1/98.
- Data was provided in multiple data sets; one for each year of the study. Data included multiple sources, see detailed conditioning sheet for more information.
- Deleted 102 records because no dates were provided.
- Deleted 194 records because no PWSIDs were provided.

Minnesota

Data received from Dick Clark and Karla Peterson.
State: Data given (12/20/97).
County: Data given (12/20/97).
PWSID: Data given (12/20/97).
Type of WSS: Data given (12/20/97).
Source Type: Data given but had to be modified in the following way: ***G*** = "GW" ***W*** = "GW" and ***S*** = "SW" (12/20/97).
PWS Name: No data given (12/20/97).
Population: Data given (12/20/97)
Sampling Point ID: ISSI generated for database purposes (2/2/98)
Sampling Point Type: Data given (12/20/97).
Sample Type: Data given (12/20/97).
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data given (12/20/97). If the reporting limit was not provided or reported as "N/A" and if the reporting code was "< x-value", the reporting limit was assumed that to be the "x-value." If the reporting limit was NA and no "< (positive result)" reported, then a reasonable assumption was made based on review of data for similar counties, PWSID, collection date, and point of contact. From June 10, 1993, the reporting limit was 1.0 ppb but prior to this period, the reporting limits were 1.0 and 5.0 ppb.
Detection: Assigned based on the reporting limit.

Conditioning Note:
- Date range: 12/92–12/97

Missouri

Data received from Darrell Osterhoudt, MO Dept of Health
State: Added "MO" to all records.
County: Data obtained from SDWIS.
PWSID: Data given (12/20/97).
Type of WSS: Data obtained from SDWIS.
Source Type: Data provided but verified with SDWIS.
PWS Name: Data given (12/20/97).
Population: Data obtained from SDWIS.
Sampling Point ID: ISSI generated for database purposes (2/2/98).

Sampling Point Type: Data determined from file titles (12/20/97).
Sample Type: Data given (12/20/97).
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data provided by data contact.
Detection: Assigned based on the reporting limit.

Conditioning Notes:
- Used ars_unc2.dbf (combination of ms_raw and ms_finished)
- Date range: 1/95–9/97
- Data contact indicated that only positive results were reported.

Montana

Data received from EPA in the earlier "11 States" data set
State: Data given (12/21/97).
County: No data given (12/21/97).
PWSID: Data given in 5-, 6-, 7- digit format. (12/21/97) used SDWIS to obtain correct PWSIDs. Some PWSID were still invalid, the data from MI, CA, ID are invalid (3/19/98).
Type of WSS: No data given from raw table (12/21/97). Stakeholders information provided this data (all data=CWS) [4/2/98]. Linked with SDWIS to determine TWSS for active systems.
Source Type: Provided converted values to IAOED standards (12/21/97).
PWS Name: Used city name, does not correlate with SDWIS system name
Population: Data given (12/21/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98).
Sampling Point Type: EPA WAM provided data (4/2/98).
Sample Type: Assumed (4/2/98).
Sample Date: No Data given.
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data obtained from state contacts. (3/20/98)
Detection: Data given (12/20/97).

Conditioning Note:
- The data contained in this data set was already modified and combined by EPA.

Nevada

Data received from SRA.
State: Added "NV" to all records.
County: Data given (12/20/97).
PWSID: Data given but added "NV" for given sysid (12/20/97).
Type of WSS: Data obtained from SDWIS.
Source Type: Data given (12/20/97).
PWS Name: Data given (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98)
Sampling Point Type: Data provided by the data contact.
Sample Type: Data provided by the data contact.
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data provided by the data contact.
Detection: Assigned based on the reporting limit.

Conditioning Note:
- Date range: 2/91–8/97.

New Hampshire

Data received from EPA in the earlier "11 States" data set

State:	Data given (12/21/97).
County:	No data given (12/21/97).
PWSID:	Data given in 5-, 6-, 7- digit format. (12/21/97) used SDWIS to obtain correct PWSIDs. Some PWSID were still invalid, the data from MI, CA, ID are invalid (3/19/98).
Type of WSS:	No data given from raw table (12/21/97). Stakeholders information provided this data (all data=CWS) [4/2/98]. Linked with SDWIS to determine TWSS for active systems.
Source Type:	Provided converted values to IAOED standards (12/21/97).
PWS Name:	Used city name, does not correlate with SDWIS system name
Population:	Data given (12/21/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	EPA WAM provided data (4/2/98).
Sample Type:	Assumed (4/2/98).
Sample Date:	No Data given.
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data obtained from state contacts. (3/20/98)
Detection:	Data given (12/20/97).

Conditioning Note:

- The data contained in this data set was already modified and combined by EPA.

New Jersey

Data received from SRA and verified with NJ contacts.

State:	Added "NJ" to all records.
County:	No data given.
PWSID:	Data given but added "NJ" to given sysid (12/20/97).
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data given but had to be modified in the following way: **G,** W="GW" **P, S,** U="SW" (12/20/97).
PWS Name:	Data given (12/20/97).
Population:	Data given (12/20/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	Data provided by the data contact.
Sample Type:	Data provided by the data contact.
Sample Date:	Data given (12/20/97).
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data given (12/20/97).
Detection:	Assigned based on the reporting limit.

Conditioning Note:

- Date range: 1/93–4/97.

New Mexico

Data received from Richard Asbury, New Mexico Environment Department

State:	Added "NM" to all records.
County:	No data given.
PWSID:	Data given.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Data given.
Population:	Data given.

Sampling Point ID: ISSI generated for database purposes
Sampling Point Type: Data provided by the data contact.
Sample Type: Data provided by the data contact.
Sample Date: Data given.
EPA Analytical Value: Data given.
Reporting Limit: Data provided by the data contact.
Detection: Assigned based on the reporting limit.

North Carolina

Data received from SRA and verified with NC contacts.
State: Added "NC" to all records.
County: County codes provided (12/20/97). Used SDWIS to fill in.
PWSID: Data given but added "NC" to given sysid (12/20/97).
Type of WSS: Data given but had to be modified in the following way: C="CWS" N="TN" P="NTNC" R="Recreation" (12/20/97).
Source Type: Data given but had to be modified in the following way: G="GW" P="SW" S="SW" W="GW" Y="GW" (12/20/97).
PWS Name: Data given (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98).
Sampling Point Type: Data provided by the data contact.
Sample Type: Data provided by the data contact.
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Were never resolved.
Detection: Assigned based on the reporting limit.

Conditioning Note:
- Date range: 4/79–4/97.

North Dakota

Data received from SRA and verified with ND contacts.
State: Added "ND" to all records.
County: No data given (12/20/97).
PWSID: Data given but added "ND" to given sysid (12/20/97).
Type of WSS: Data obtained from SDWIS.
Source Type: Data given (12/20/97).
PWS Name: Data given (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98).
Sampling Point Type: Data provided by the data contact.
Sample Type: Assumption made the data contact.
Sample Date: Data given (12/20/97).
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data provided by the data contact.
Detection: Data given (12/20/97).

Conditioning Note:
- Date range: 1/93–10/96.

Ohio

Data received from EPA in the earlier "11 States" data set

State:	Data given (12/21/97).
County:	No data given (12/21/97).
PWSID:	Data given in 5-, 6-, 7- digit format. (12/21/97) used SDWIS to obtain correct PWSIDs. Some PWSID were still invalid, the data from MI, CA, ID are invalid (3/19/98).
Type of WSS:	No data given from raw table (12/21/97). Stakeholders information provided this data (all data=CWS) [4/2/98]. Linked with SDWIS to determine TWSS for active systems.
Source Type:	Provided converted values to IAOED standards (12/21/97).
PWS Name:	Used city name, does not correlate with SDWIS system name
Population:	Data given (12/21/97).
Sampling Point ID:	ISSI generated for database purposes (2/2/98).
Sampling Point Type:	EPA WAM provided data (4/2/98).
Sample Type:	Assumed (4/2/98).
Sample Date:	No Data given.
EPA Analytical Value:	Data given (12/20/97).
Reporting Limit:	Data obtained from state contacts. (3/20/98)
Detection:	Data given (12/20/97).

Conditioning Note:

- The data contained in this data set was already modified and combined by EPA.

Oklahoma

Data received from Tom Poeten, EPA Region 6.

State:	ISSI generated for database purposes.
County:	Provided in database.
PWSID:	Database provided abbreviated PWSID numbers, which were converted to complete PWSID numbers in accordance with directions from Tom Poeten.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Provided in database.
Population:	Data obtained from SDWIS.
Sampling Point ID:	Provided in database.
Sampling Point Type:	Provided in database.
Sample Type:	Provided in database.
Sample Date:	Provided in database.
EPA Analytical Value:	Provided in database.
Reporting Limit:	Provided in database.
Detection:	Provided in database.

Conditioning Note:

- Deleted 4 samples with unusual dates (years reported at 05).

Oregon

Data received from Ed Thomas EPA, verified with Patrick Meyer, OR

State:	ISSI generated for database purposes.
County:	Provided in database.
PWSID:	Provided in database.
Type of WSS:	Data obtained from SDWIS.
Source Type:	Data obtained from SDWIS.
PWS Name:	Provided in database.
Population:	Data obtained from SDWIS.

Sampling Point ID: Provided in database.
Sampling Point Type: Provided in database.
Sample Type: Provided in database.
Sample Date: Provided in database.
EPA Analytical Value: Provided in database.
Reporting Limit: Not provided in earlier samples (ND = 0). Assumed to be 5 in 1993, and 1 in 1994 and 1995 based on frequency analysis. 1994 and later samples also included an increasing number of samples with reporting limits.
Detection: Provided in database.

Texas

Data received from SRA and verified with TX contacts.
State: Added "TX" to all records.
County: No data given (12/20/97).
PWSID: Data give but added "TX" to given sysids.
Type of WSS: Data obtained from SDWIS.
Source Type: Data given but had to modified in the following way: **S,** P="SW"; **G,** W="GW"; Y – blank…(SDWIS).
PWS Name: Data give (12/20/97).
Population: Data given (12/20/97).
Sampling Point ID: ISSI generated for database purposes (2/2/98)
Sampling Point Type: Data provided by the data contact.
Sample Type: Data provided by the data contact.
Sample Date: No data given.
EPA Analytical Value: Data given (12/20/97).
Reporting Limit: Data provided by the data contact.
Detection: Based on the reporting limit

Conditioning Notes:
- Date range: 3/92–12/96.
- Changed "0" values to non-detects at the reporting limit; there were 51 of these cases. Because the reporting limit was determined by the date, and some records had no date, they had to be removed, there were 13 of them.

Utah

Data received from Larry Scanlon UT Dept of Env. Quality
State: Added "UT" to all records.
County: Data obtained from SDWIS.
PWSID: Data obtained from SDWIS.
Type of WSS: Data obtained from SDWIS.
Source Type: Data obtained from SDWIS.
Purchased: Data obtained from SDWIS.
PWS Name: Data given (12/20/97 and 3/99).
Population: Data obtained from SDWIS.
Sampling Point ID: ISSI generated for database purposes (2/2/98)
Sampling Point Type: Data given (12/20/97 and 3/99).
Sample Type: Data given (12/20/97 and 3/99).
Sample Date: Data given (12/20/97 and 3/99).
EPA Analytical Value: Data given (12/20/97 and 3/99).
Reporting Limit: Data given. Used less than values as reporting limits.
Detection: Less than values were classified as nondetects.

Conditioning Notes:

- Used data from nciver2.xls file.
- Data from utahas96.wq2 seems to be the data from two of the SRA Utah raw files (utahpv and utahpw).
- Data updated and additional samples added with results received from Larry Scanlon (via EPA WAM) in March 1999.
- Date range: 1/78–3/99.
- Deleted S5921 (zero result reported for detection).

This page intentionally left blank

Appendix D-2
AOED Database Specifications

This page intentionally left blank

Contents of GRAND

The CONTENTS Procedure

Data Set Name:	FINAL.GRAND	Observations:	131383
Member Type:	DATA	Variables:	15
Engine:	V8	Indexes:	0
Created:	11:18 Friday, December 22, 2000	Observation Length:	144
Last Modified:	11:18 Friday, December 22, 2000	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	YES
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	12288
Number of Data Set Pages:	1546
First Data Page:	1
Max Obs per Page:	85
Obs in First Data Page:	65
Number of Data Set Repairs:	0
File Name:	grand.sas7bdat
Release Created:	8.0000M0
Host Created:	WIN_95

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Informat	Label
9	CNTY	Char	30	105	$30.	$30.	County
12	COLLDATE	Num	8	8	DATE9.	YYMMDD8.	Collection Date (SASdate)
15	DETECT	Char	1	137	$1.	$1.	Detect Flag (N,D)
2	FILL	Char	1	41			Estimated Reporting Limit? (Y=yes, N=no)
7	POP	Num	8	0	11.	11.	Population
1	PWSID	Char	9	32			PWS ID
6	PWSNAME	Char	50	53	$50.	$50.	PWS Name
13	RESULT	Num	8	16	20.5	20.5	Concentration (ug/L)
14	RPTLIMIT	Num	8	24	20.5	20.5	Reporting Limit (ug/L)
10	SAMPTTYP	Char	1	135	$1.	$1.	Sample Type (R=raw, F=finished)
8	STATE	Char	2	103	$2.	$2.	State Abbreviation
5	STYPE	Char	2	51	$2.	$2.	PWS Source Type (GW,SW)
3	ST_GROUP	Char	3	42			State Group
11	TOT_DISS	Char	1	136	$1.	$1.	Total or dissolved (T, D)
4	TWSS	Char	6	45	$6.	$6.	Type Water System (CWS,NTNCWS)

-----Sort Information-----

Sortedby:	ST_GROUP PWSID COLLDATE
Validated:	YES
Character Set:	ANSI

Contents of INTRA

The CONTENTS Procedure

Data Set Name:	FINAL.INTRA	Observations:	88855
Member Type:	DATA	Variables:	17
Engine:	V8	Indexes:	0
Created:	11:18 Friday, December 22, 2000	Observation Length:	160
Last Modified:	11:18 Friday, December 22, 2000	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	YES
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	16384
Number of Data Set Pages:	872
First Data Page:	1
Max Obs per Page:	102
Obs in First Data Page:	83
Number of Data Set Repairs:	0
File Name:	intra.sas7bdat
Release Created:	8.0000M0
Host Created:	WIN_95

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Informat	Label
9	CNTY	Char	30	105	$30.	$30.	County
12	COLLDATE	Num	8	8	DATE9.	YYMMDD8.	Collection Date (SASdate)
15	DETECT	Char	1	137	$1.	$1.	Detect Flag (N,D)
2	FILL	Char	1	41			Estimated Reporting Limit? (Y=yes, N=no)
7	POP	Num	8	0	11.	11.	Population
1	PWSID	Char	9	32			PWS ID
6	PWSNAME	Char	50	53	$50.	$50.	PWS Name
13	RESULT	Num	8	16	20.5	20.5	Concentration (ug/L)
14	RPTLIMIT	Num	8	24	20.5	20.5	Reporting Limit (ug/L)
10	SAMPTTYP	Char	1	135	$1.	$1.	Sample Type (R=raw, F=finished)
16	SRC_ID	Char	20	138			Entry Point
17	SRC_TYP	Char	2	158			Entry Point Type
8	STATE	Char	2	103	$2.	$2.	State Abbreviation
5	STYPE	Char	2	51	$2.	$2.	PWS Source Type (GW,SW)
3	ST_GROUP	Char	3	42			State Group
11	TOT_DISS	Char	1	136	$1.	$1.	Total or dissolved (T, D)
4	TWSS	Char	6	45	$6.	$6.	Type Water System (CWS,NTNCWS)

-----Sort Information-----

Sortedby:	ST_GROUP PWSID COLLDATE
Validated:	YES
Character Set:	ANSI

Appendix D-3
Initial Data Conditioning Process

This page intentionally left blank

DATABASE DEVELOPMENT AND DATA CONDITIONING

This appendix summarizes the initial development of the Arsenic Occurrence and Exposure Database (AOED). Further modifications to the database are described in chapter 4, section 4.1.3. These modifications include the replacement of some state data with updated data and the use of updated SDWIS information.

Database Design

A two step process was used to identify the database design that is necessary for developing arsenic occurrence and exposure projections. First, data elements were identified by analyzing similar data models used for arsenic occurrence and exposure projections. Second, a database was designed to support estimation of arsenic occurrence and exposure, to accommodate the data elements identified in step one, and to maximize functionality.

Data Element Identification

ISSI identified data elements for AOED from USEPA's emerging National Contaminant Occurrence Database (NCOD) and SAIC's previous arsenic occurrence project. The data elements chosen for inclusion are the minimum elements required to estimate arsenic occurrence and exposure, as well as additional elements chosen to support the Analysis Plan and the Arsenic Occurrence and Exposure Report. Four different categories of data related to arsenic occurrence and exposure were established to identify data elements with common characteristics, including Location Information, PWS Information, Sample Information, and Result Information. The Location Information category contains information on the USEPA Region, State, and County. The PWS Information category contains the name, address, PWSID #, purchase category, type of water system supply (Community Water System (CWS) or Non-Transient Non-Community Water System (NT NCWS)), population, PWS latitude and longitude, and source type (ground water or surface water). The Sample Information category table contains sampling point type (finished or raw), sample ID, sample latitude and longitude, sample type (total or dissolved), and sample collection date. Finally, the Contaminant Information category contains the contaminant, USEPA analytical value, unit of measure, reporting limit, detection, and speciation.

Database Structure

The Arsenic Occurrence and Exposure Database (AOED) was designed to support development of estimates of arsenic occurrence and exposure. The four categories (location information, PWS information, sample information, and result information) summarized above correspond to tables in the relational database. The tables are related in a one to many relationship, starting with the location table and ending with the contaminant table. Data elements identified for each category correspond to a column in the related table.

In addition to the data elements identified above, three source tables were included to identify the source of each data point. For example, for data elements obtained from the NIRS data set, the corresponding data elements in the source tables were populated with the code 'NR'. Each of the source tables is related to its parent data table in a one-to-one relationship.

Database Review Process

The database review process evaluated the potential value of the data for use in projecting the national occurrence and exposure estimates for arsenic. Raw data sets (RDSs) that are suitable for use in the arsenic projections were identified as a result of this process. The database review process took place in three stages.

The first stage was to identify and evaluate the characteristics of the RDSs. An interaction analysis was conducted, where the data elements in each RDS was compared with the data elements specified in the AOED requirements. An interaction analysis identifies the relationship between data elements in two distinct data sets. Several other analyses were conducted to identify the critical data elements within each RDS and to evaluate the RDS data quality.

The second stage of the data review process was conditioning. The results of the interaction analysis were used to guide the population of a temporary database with suitable RDS data formatted to fit the specifications of AOED.

The final stage was data gathering and data set decision. Data gathering consisted of identification of data gaps and populating them with acceptable data. Data set decision involved the evaluation of a data set to determine if a sufficient amount of information was present to support AOED. Data sets that did not contain sufficient information to support arsenic occurrence and exposure estimates were removed from the development process.

Raw Data Set Analyses

The RDSs consisted of a wide variety of data elements and formats. RDS analyses were conducted to identify:

1) data quality and format,
2) critical data elements within each RDS, and
3) the relationship between the RDS and AOED data elements

The first type of analysis examined the characteristics of the data elements. These characteristics varied by data element and data source. For example, for the Source Type data element, one RDS might use ‘G’ to define a groundwater source, where another RDS might use ‘GW,’ while a third might use a numerical code such as 3.

The second type of analysis identified the critical data elements within each RDS. These critical data elements were necessary to project arsenic occurrence and exposure. The critical data elements identified were Source Type, Type of Water System Supply, PWS ID#, Reporting Limit, Sample Collection Date, and Detection. Analyses also included a report on population distribution within each of the RDS, the spatial and temporal coverage provided by the RDS, and miscellaneous other analyses. These analyses provided a preliminary overview of the data quality and representativeness. The results of these analyses were used to present interim status reports to USEPA, identify data gaps, and provide a basis for the project direction.

The third type of analysis identified the relationship between the data elements present in the RDS and those specified for AOED. Four types of relationships were identified between RDS and AOED data elements. In the first type of relationship, a data element found in the RDS has a definition which matches the definition of the AOED data element. This is known as a correlated element. The second type of relationship is a calculated value, which occurs when one or more data elements in the RDS can be manipulated to calculate the value of the AOED data element. The third relationship is a logical inference, where the value for the AOED data element can be logically inferred from background information about the RDS. The final relationship, no correlation, exists when no data element in the RDS correlates with the AOED data element.

Data Conditioning

The data from the RDS were transferred to spreadsheets to facilitate further analyses. Each spreadsheet was associated with one RDS and contained the data elements for AOED. These spreadsheets were used to compile all of the raw data into a single uniform format that, made it easy to perform data manipulation. Collectively, these "data sheets" comprised IAOED (*Intermediate* Arsenic Occurrence and Exposure Database). IAOED was a necessary step to properly format the data for input into AOED, the relational database.

IAOED played an important role in the identification of data gaps and RDS quality concerns, and supported initial data analyses, the interaction analysis, and data conditioning. Data conditioning involved using the results of the interaction analysis to guide the transfer of data element values into IAOED according to the conditioning steps indicated below:

Case 1. Correlated Elements - The data element values were copied directly from the RDS to IAOED. No modifications were necessary to the data, as the data element definitions matched closely. A data element was found in the RDS whose definitions correlate closely with the definition of the AOED data element. The value of the RDS data element can be used directly as the value of the AOED data element. For example, the AOED data element **Sample Collection Date** conforms directly to the SRA-Texas data element **SAMP_DATE**.

Case 2. Calculated Value - The data element values for IAOED were determined, using an algorithm, from the RDS value to the properly formatted AOED value. To illustrate, the AOED data element **Sampling Point Type** can be derived by converting the data element **ANALYTNAME** from the Missouri RDS according to the following algorithm:

If ANALYTNAME = "Arsenic, Total", then code as "T"
If ANALYTNAME = "Arsenic, Dissolved", then code as"D"
If ANALYTNAME = " ", then leave field blank.

Case 3. Logical Inference from RDS Characteristics - The data element value for IAOED was inferred from the background information known about the RDS source. In most cases, the same value was used for all records in a single IAOED data sheet. For example the AOED data element **USEPA Region** is not present in the SRA-Alaska, but Alaskan water supplies all occur in USEPA Region 10, so the value for **USEPA Region** can be populated with "10".

Case 4. No correlation - The data element value for IAOED was not populated with data from the RDS. Efforts were made to contact the source of the data to pursue possible values for the data element.

IAOED data conditioning involved population of the source data tables with information to track the source of each data point, and documentation of the conditioning process to track quality assurance and quality control (QA/QC) issues. Data conditioning also involved generation of artificial Loc ID#, Sampling Point ID#, and Result ID# if these numbers were not provided in the State databases in anticipation of the transfer of data from the loosely structured "data sheet" format of IAOED to the controlled relational structure of AOED. For the eight States with intra-system data, Sampling Point ID or Loc ID were provided so and we could track samples that were collected from individual POE.

Data Gathering and Decision

After preliminary data conditioning, and identification of data gaps, efforts were made to fill in the data gaps with accurate data. Representatives from the agency or organization which provided the RDSs were contacted to obtain the missing information. For several of the RDSs, information could not be obtained for critical data elements. These RDSs should not be used in the arsenic occurrence projections due to insufficient information, rounding of results, duplication of results in another data set or unknown manipulation of data.