Supporting Information

Title: Using compliance data to understand uncertainty in drinking water lead levels in southwestern Pennsylvania

Authors: Sara E. Schwetschenau; Mitchell J. Small; Jeanne M. VanBriesen

Number of Pages: 35

Number of Figures: 35

Number of Tables: 4

1 Overview of Lead and Copper Rule sample site selection requirements

The Lead and Copper Rule sets specific requirements for how sites are selected for compliance sampling. The current requirements are outlined in Figure SI.1

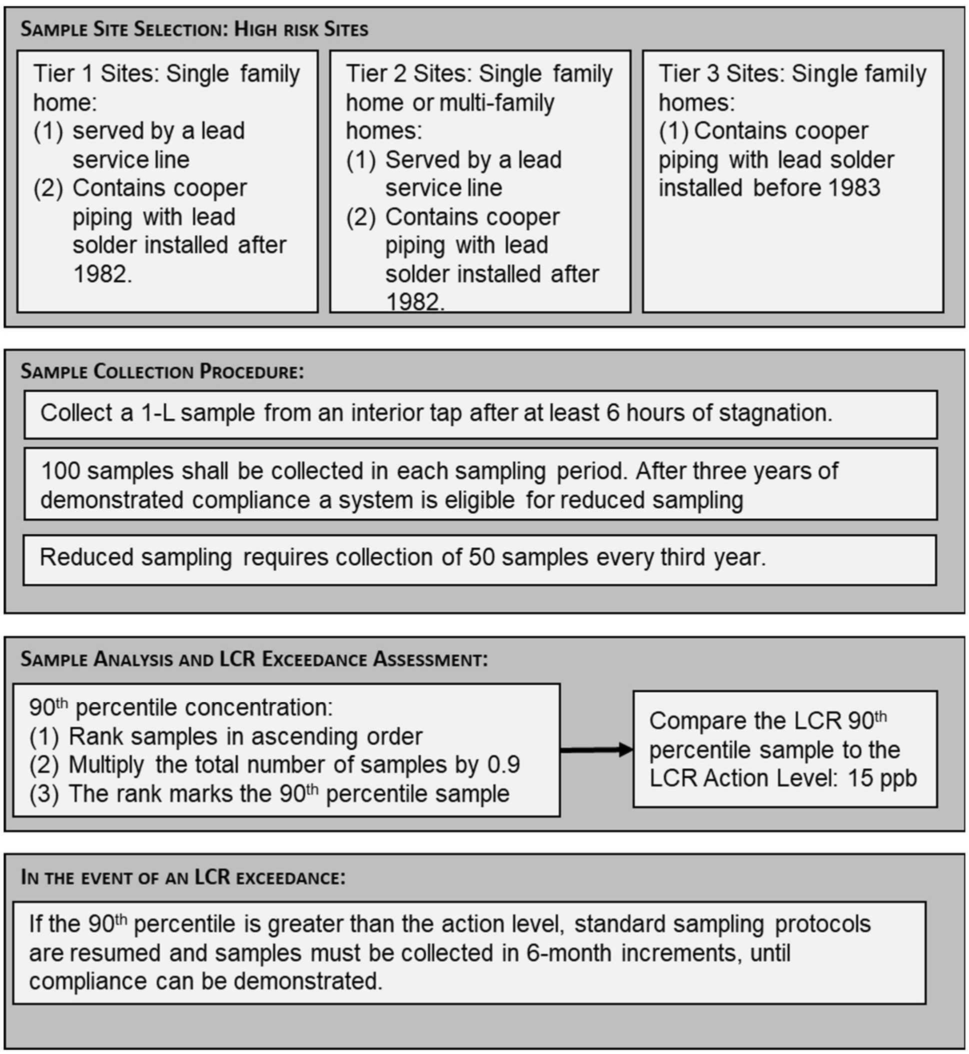

Figure SI.1. Lead and Copper Rule site selection protocol

Sites sampled to assess compliance with the LCR are intended to be high-risk for lead release. Sites are organized into one of three tiers. Tier 1 sites are considered at highest risk and included single family homes with a lead service line or lead solder installed after 1982, as newer lead solder was assumed to be more likely to release lead than older solder [1]. The proposed LCR revision calls for prioritizing sampling from lead service lines[2]. If a sufficient number of Tier 1 sites cannot be identified, Tier 2 sites are sampled. Tier 2 sites include all the same requirements as Tier 1 sites, but are multi-family residences. Single family residences were identified as preferred over multi-family homes as less variation in the interior plumbing configuration is expected; samples from single family homes are assumed to be more comparable. If a utility cannot identify a sufficient number of Tier 1 or Tier 2 sites, then Tier 3 sites are used. These sites include locations with lead solder installed prior to 1983 and, given the age of the plumbing (greater than five years old at the time of the LCR writing), are thought to release less lead than newer installations of lead solder. Classification of sites based on the tiered system are based on a materials assessment completed by the utility through review of building records, renovation permits, and other housing documentation.

The original version of the LCR required the samples be selected with priority given to locations at the end of the distribution system [1]. Lead levels are expected to be higher at the ends of a distribution system due to the long detention time; any corrosion control chemicals added at the entry point to the distribution system may have been consumed prior to reaching that point. This requirement was not included in the final version of the regulation [1].

After identification of a sampling pool of sites that meets the criteria of the tiered system, samples are collected (second dark gray box in Figure SI.1). Samples are required to be a first-draw one-liter sample taken at an indoor tap after a minimum of six hours of stagnation. These samples are typically collected by the resident and returned to the water utility by mail for testing. While this setup is likely the least intrusive for the resident, as the water utility does not require access to their home, it is often not possible to verify if all samples were collected consistent with the protocol, increasing the likelihood of sampling error. Initial versions of the LCR allowed for collection of service line samples, where the line would be flushed after stagnation until the sample collected was thought to have been stagnant in the service line, rather than in interior plumbing as is the case for the first-draw sample. Current research [3–5] where profile samples were collected, shows that the highest concentrations of lead come from lead service lines rather than the interior plumbing or fixtures sampled with a first-draw sample alone. As such a first draw sample can underestimate the lead exposure associated with a specific site.

A sample size of 100 homes is initially required, with the option to reduce the sample size to 50 homes after compliance has been demonstrated for three consecutive years. Once the annual sampling is completed, the 90th percentile concentration is determined (see third dark gray box in Figure SI.1). Figure SI.1 summarizes the steps required to determine the 90th percentile for a specific sample set. If the calculated 90th percentile is greater than the action level of 15 ppb, then the utility has triggered an exceedance of the LCR and additional steps must be completed. These include increasing the sampling frequency to every 6 months, deployment or adjustment of corrosion control treatment and removal of 7-percent of remaining lead service lines [1].

Sample sets analyzed for this work were collected based on the requirements described here. Most sample sets were taken under reduced sampling requirements, while others were collected following an exceedance of the LCR and reflect the increased sampling frequency and number of samples described above.

2 Descriptive statistics

To inform subsequent model selection, descriptive statistics were performed and reviewed. Figure SI.2 shows a scatterplot of all sample sets grouped by water utility and shown in order of sample year. These raw data do not include the imputed sample estimates; all samples recorded as below the reporting limit are plotted at zero (matching how they are reported in the DWRS online reporting system[6]). Samples are binned and colored based on the concentration of the sample to show how the number of samples is split across the results range. The highest sample concentrations recorded are over 300 ppb for system A. Three samples greater than 150 ppb were recorded between systems A and C. Even for these utilities with the highest recorded sample concentrations, the majority of the samples collected are at or near the reporting limit (1 ppb). Table 1 shows the number of samples in each sample set and the number of samples below the reporting limit and over the action limit.

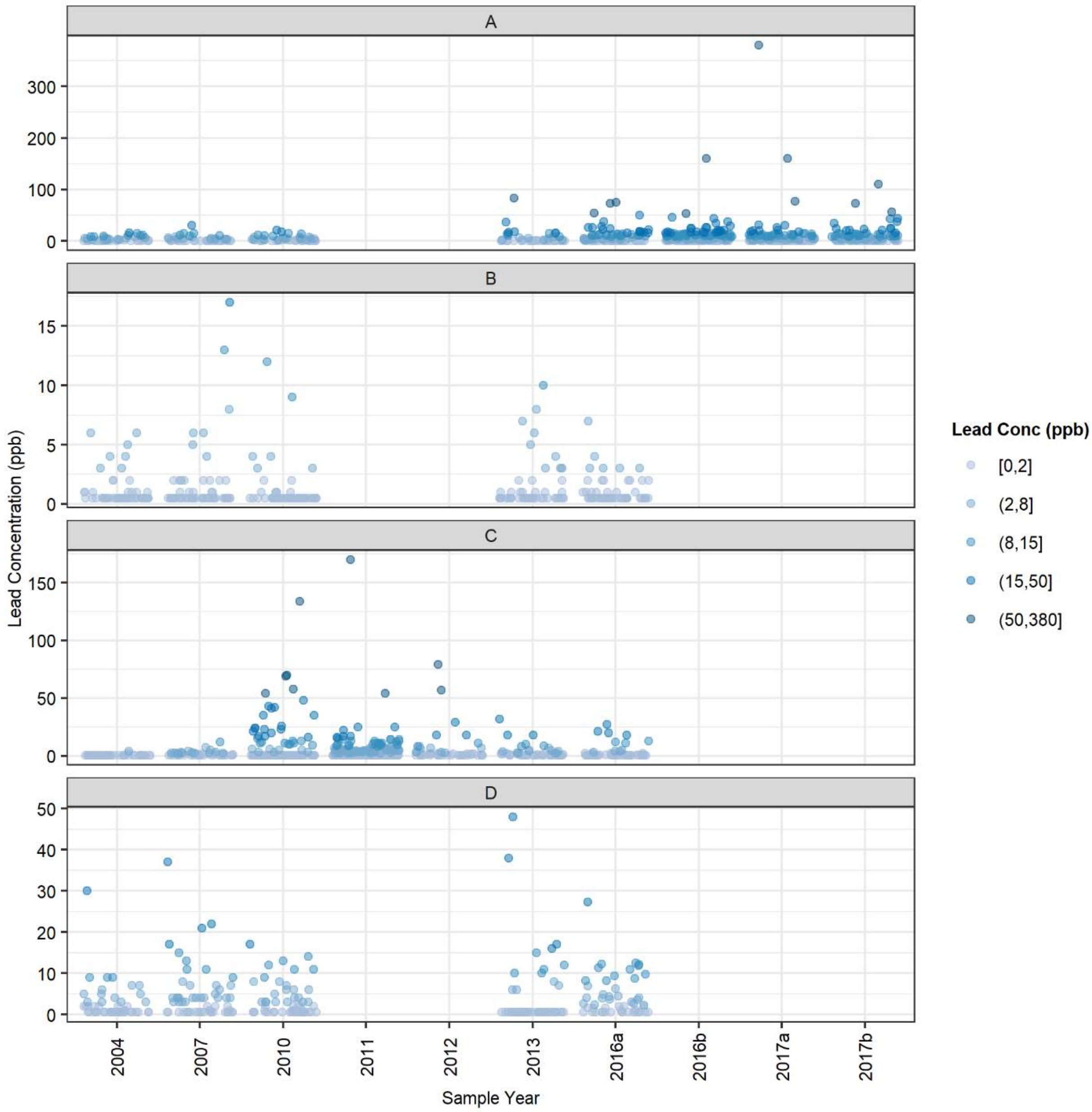

Figure SI.2. Scatterplot of LCR lead data for utilities A through D from 2004 through 2017. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year.

Figure SI.3 shows the imputed data sets as a boxplot. Red dots show the 90th percentile of the distribution and the red line is the LCR action level of 15 ppb. The blue dotted line is the reporting limit. The median of the distribution is shown by the black line with the edges of the box representing the 25^{th} to 75^{th} percentile range. Black dots represent outliers greater than 1.5 times the interquartile range denoted by the whiskers on the plot. In contrast with Figure 1, Figure SI.3 uses an arithmetic scale (as opposed to log-transformed in Figure 1) and allows for improved interpretation of the extremes of the distribution system and comparison between the extreme values and the median or 75^{th} percentile for each distribution. Less variation is observed between the medians or 75^{th} percentile value among sample sets than between extreme values. For example, Utility C has a median of 1.3 ppb and 90th percentile value of 3.2 ppb in 2007 compared to a median of 2.7 ppb and a 90th percentile value of 30.5 ppb in 2010 (values shown in Table 1). Utility A shows a greater range and a greater number of outlier samples than all other years across other systems, with the exception of Utility C in 2010. Utility B has very few outliers and nearly all samples collected are below the action level of the 15 ppb. Utility D shows a 90^{th} percentile value near the action level, but no samples greater than 50 ppb, and a limited number of samples greater than the action level, when compared to Utilities A and C. Table 1 shows the exact number of samples over the action level for each sample set. All of these trends indicate a strong positive skew and a long right tail across most of these distributions; the length of the tail and the range over which outliers extend, varies among sample sets.

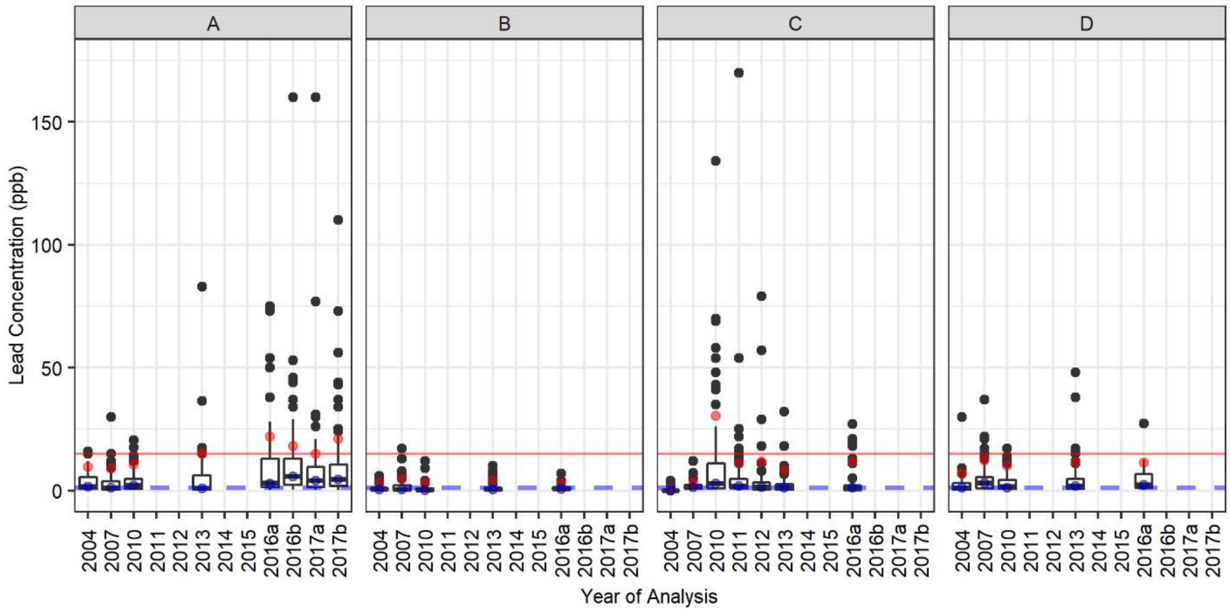

Figure SI.3. Boxplot of imputed data for utilities A through D from 2004 through 2017. Years designated as "a" or "b" indicate two sample periods were conducted that year. "a" represents samples collected between January and June and "b" represents samples collected between July and December of the indicated year.

While Kaplan-Meier methods were used here to account for samples below the reporting limit when estimating the summary statistics, substitution is a common alternative approach. To evaluate the effect of substitution methods over Kaplan-Meier, Table SI.1 shows the variation in the sample mean for each sample set for different methods of handling the samples below the reporting limit. Methods compared include: Kaplan-Meier, statistics applied to the imputed sample set (generated using Kaplan Meier and regression on order statistics, and substitution (three substituted values were used 0, 0.5 and 1) prior to calculation of statistics. The maximum percent change in the calculated mean between Kaplan Meier estimates and all other methods is shown in the right most column for comparison. The closer the mean is to the reporting limit the greater the variation in the estimates; these are highest across all years for Utility B. Utility C in

2004 has a very high variation because 48 of the 50 samples collected were all below the reporting limit. In these cases, any imputation method may result in greater uncertainty in the mean, but for these years, there is also likely public health risk associated with elevated lead levels. The difference in the mean statistic is highest when Kaplan Meier estimates are compared to substitution with a value of 0. Using 0 (the bottom of the range) is not a conservative estimate and there is no information to suggest that 0 is any more likely than another value within the range of 0 to 1. For systems where the 90th percentile is at or near the regulatory limit, and consideration of the uncertainty bounds important, the effect of using Kaplan Meier imputation is small and unlikely to have affected estimates of the 90th percentile (e.g., Utility A in 2016-2017 and Utility C in 2010).

Table SI.1. Comparison of sample statistics using imputed values versus substitution to estimate values for samples recorded as below the reporting limit

		Using Kaplan Meier[3]	Using imputed values[4]	substituted BRL data as 0[5]	substituted BRL data as 0.5[6]	substituted BRL data as 1[7]	Max %-change from Kaplan Meier[8]
Year[1]	Utility ID[2]	Mean	Mean	Mean	Mean	Mean	
2004	A	3.74	3.54	3.346	3.526	3.706	10.5%
2007		3.34	3.34	3.12	3.37	3.62	8.4%
2010		3.99	3.83	3.7	3.8	3.9	7.3%
2013		5.84	5.84	5.628	5.888	6.148	5.3%
2016a		9.36	9.02	8.415	8.64	8.865	10.1%
2016b		9.66	9.52	8.988	9.148	9.309	7.0%
2017a		10.96	10.58	10.202	10.381	10.561	6.9%
2017b		9.57	9.29	8.897	9.058	9.219	7.0%
2004	B	1.07	1.07	0.88	1.21	1.54	43.9%
2007		1.61	1.61	1.439	1.754	2.07	28.6%
2010		1.01	1.01	0.88	1.25	1.62	60.4%
2013		1.41	1.41	1.24	1.55	1.86	31.9%
2016		1.1	1.1	0.846	1.154	1.462	32.9%
2004	C	0.2	0.2	0.12	0.6	1.08	440.0%
2007		1.99	1.79	1.58	1.767	1.953	20.6%
2010		10.35	10.348	9.302	9.623	9.943	10.1%
2011		4.94	4.776	4.615	4.775	4.935	6.6%
2012		6.06	5.765	5.664	5.864	6.064	6.5%
2013		3.44	3.144	2.998	3.218	3.438	12.8%
2016		3.46	3.078	2.981	3.222	3.463	13.8%
2004	D	2.75	2.75	2.48	2.74	3	9.8%
2007		5.09	4.995	4.745	4.918	5.091	6.8%
2010		3.57	3.379	3.179	3.375	3.571	11.0%
2013		4.88	4.88	3.709	4.091	4.473	24.0%
2016		4.52	4.423	4.204	4.36	4.517	7.0%

Notes:

1. Year in which the samples were collected for each utility. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year.
2. Utility ID – unique identifier for the water system analyzed.
3. Mean of each sample set computed using Kaplan Meier methods.
4. Mean of each sample set computed after imputing estimates below the reporting limit.
5. Mean of each samples set computed after substituting a value of zero for each sample below the reporting limit.
6. Mean of each samples set computed after substituting a value of 0.5 ppb, or one-half the report limit, for each sample below the reporting limit.
7. Mean of each samples set computed after substituting a value of 1 ppb, or the report limit, for each sample below the reporting limit.
8. The percent change in the mean between each method and Kaplan Meier is calculated and the maximum value shown.

3 Model fitting and selection

The lognormal transformation is common for environmental data sets and used here. The transformed datasets is then fit to the normal and student’s t-distribution and the model fits evaluated. The mathematical form for the lognormal transformation is shown below.

$$y_i = \ln\left(x_i\right)$$

The mathematical form of the probability density function (pdf) is shown below for the normal and student’s t-distribution, respectively.

$$f\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-\left(y_i - \mu\right)^2}{2\sigma^2}\right\}$$

$$f\left(y_i\right) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{y_i^{\,2}}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} \textit{ where } \upsilon \textit{ is the degrees of freedom}$$

Figure SI.4 through Figure SI.28 show the visual fit of each lognormally transformed sample set compared to the normal and student’s t-distribution. The two panels on the left are the quantile-

quantile plots comparing the fitted distribution to the sample set and on the right is a comparison of the empirical CDFs comparing simulated predicted samples from each fitted distribution to the original imputed sample set. For the quantile-quantile plots, alignment of the dots with the black diagonal line indicates a consistent and good fit. Where the dots deviate from the line, there is an indication that the range of the sample set is not well represented by the selected model. For example, in Figure SI.4 for the lognormal distribution (top left), values in the right tail are below the diagonal, indicating that the fitted distribution may be overpredicting estimates in this range. The rest of the distribution is a decent approximation, but does deviate in places along the distribution. While this could just be a factor of natural deviation and small sample sets, it could also indicate that the data are not well represented by a single distribution, and mixture model fits[9] should be evaluated if an improved fit is desired. In Figure SI.4 the same trends seen in the quantile-quantile plot are reinforced through evaluation of the CDFs, with the student's T-distribution model showing greater deviation from the sample than the lognormal model fit. This same type of visual evaluation was repeated for all sample sets shown in the subsequent figures. Based on this assessment the lognormal distribution was considered to be an acceptable model fit for all sample sets.

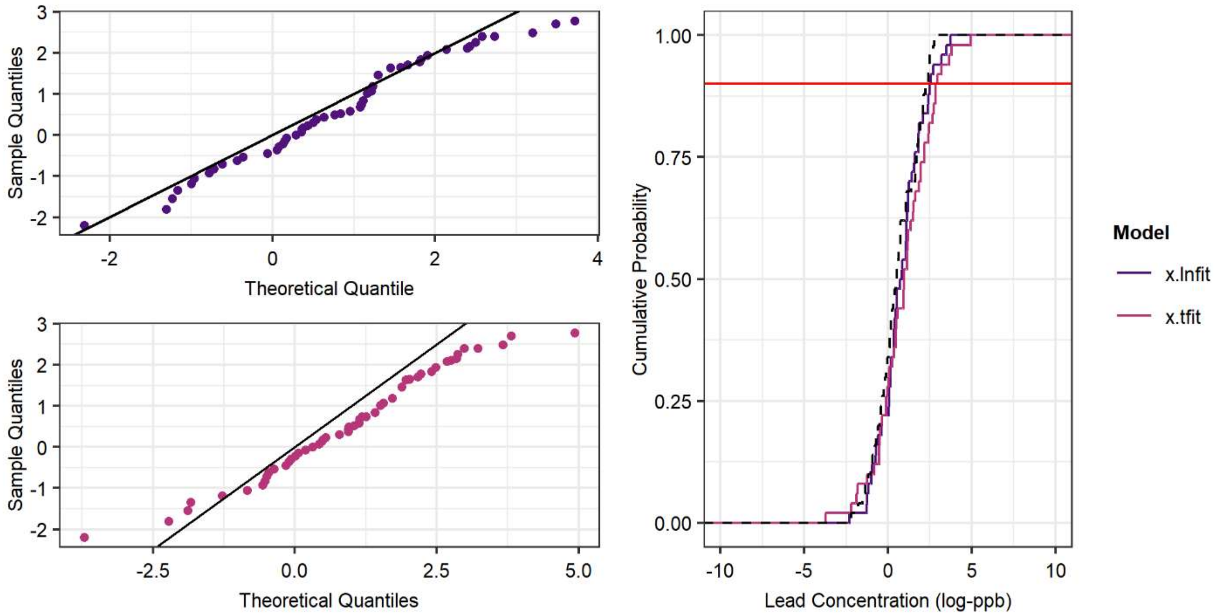

Figure SI.4. Utility A 2004 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

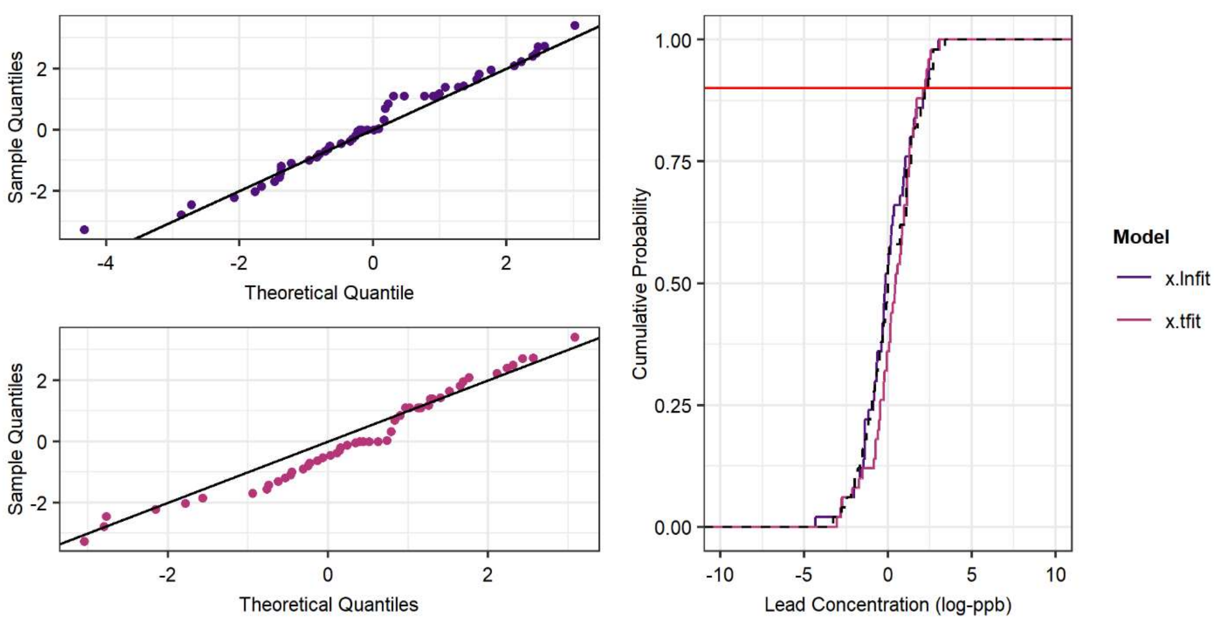

Figure SI.5. Utility A 2007 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

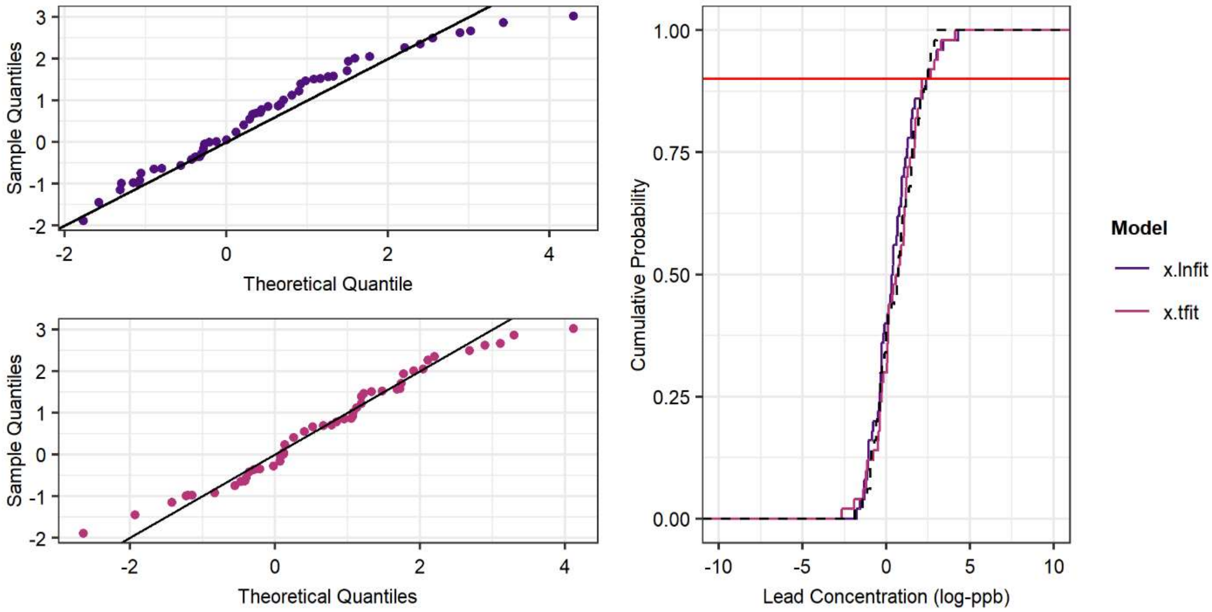

Figure SI.6. Utility A 2010 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

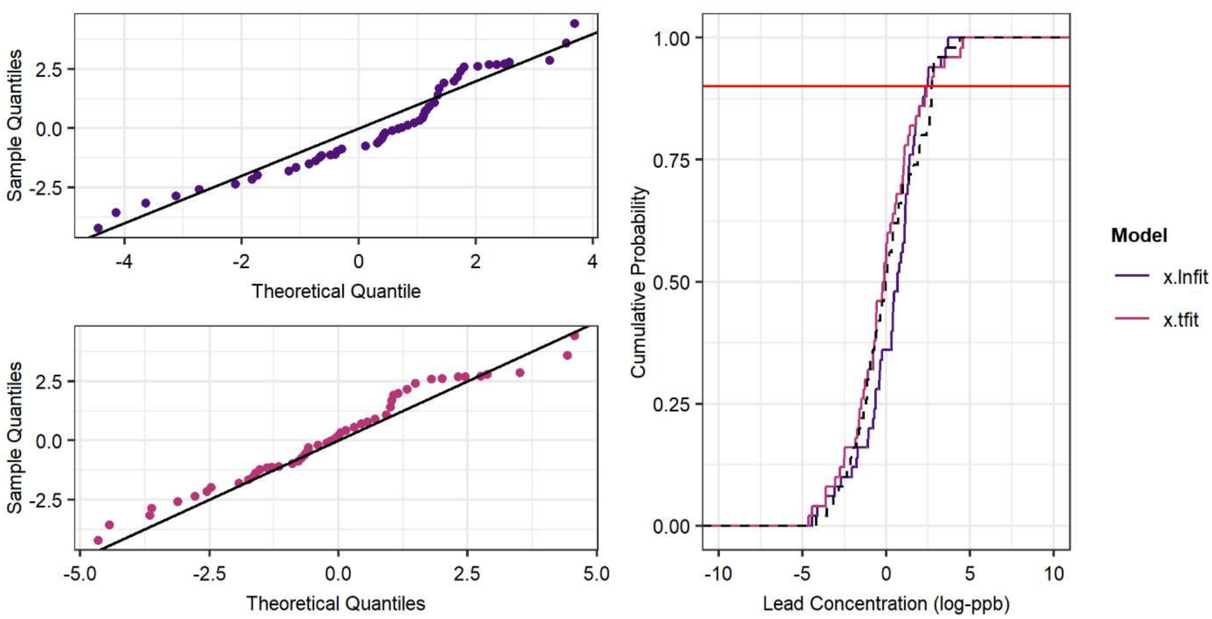

Figure SI.7. Utility A 2013 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

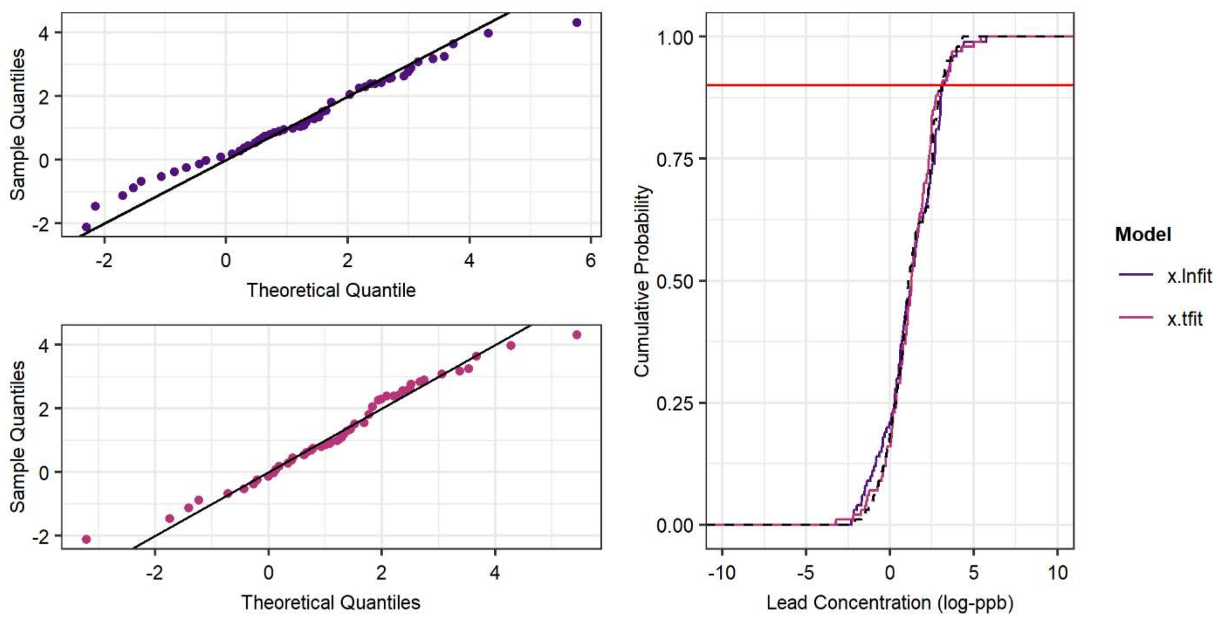

Figure SI.8. Utility A 2016a (January to June) – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

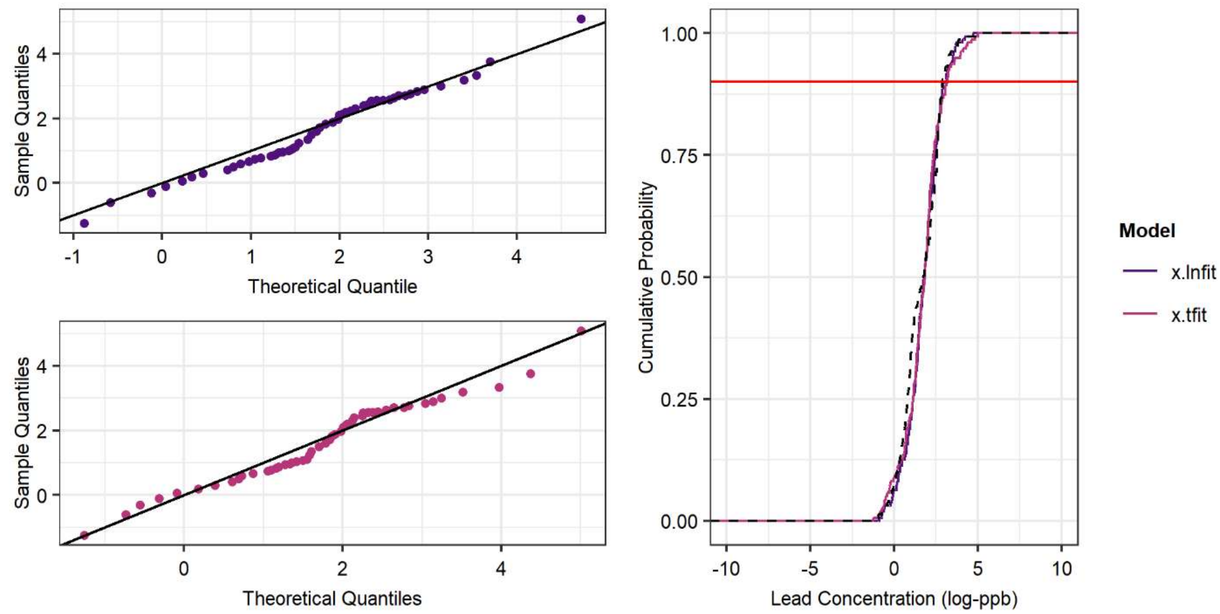

Figure SI.9. Utility A 2016b (July to December) – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

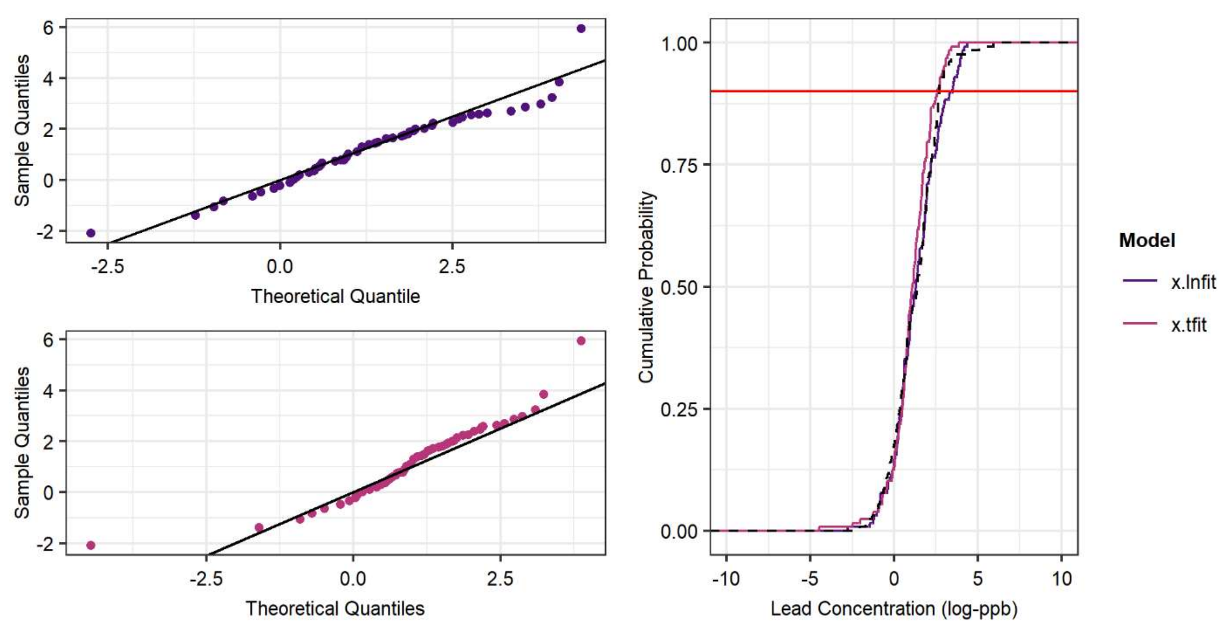

Figure SI.10. Utility A 2017a (January to June)– Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

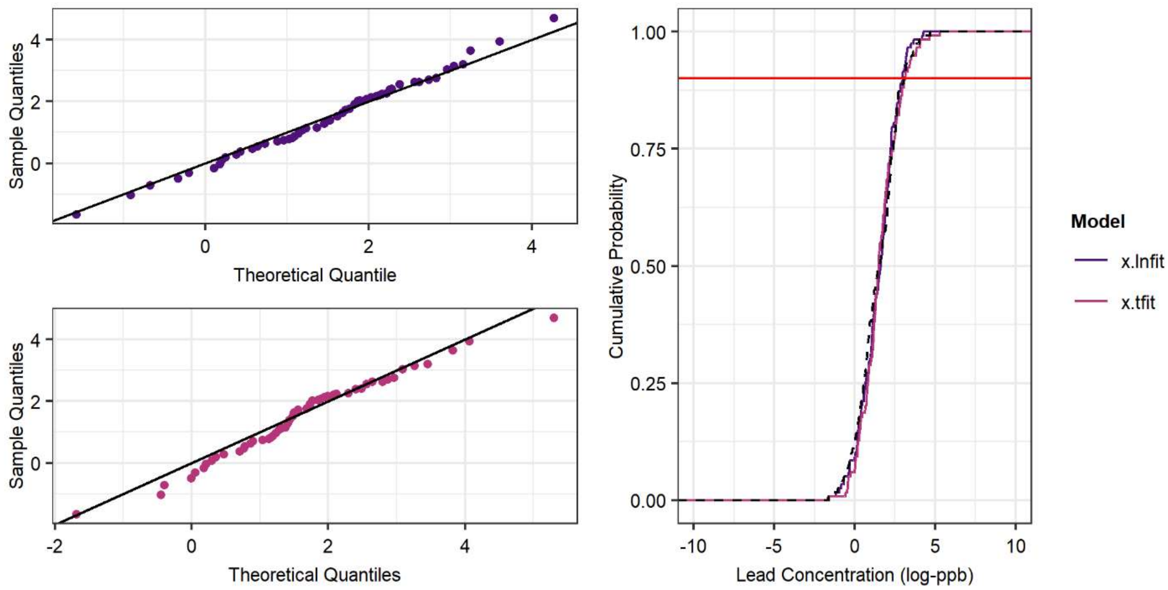

Figure SI.11. Utility A 2017b (July to December) – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

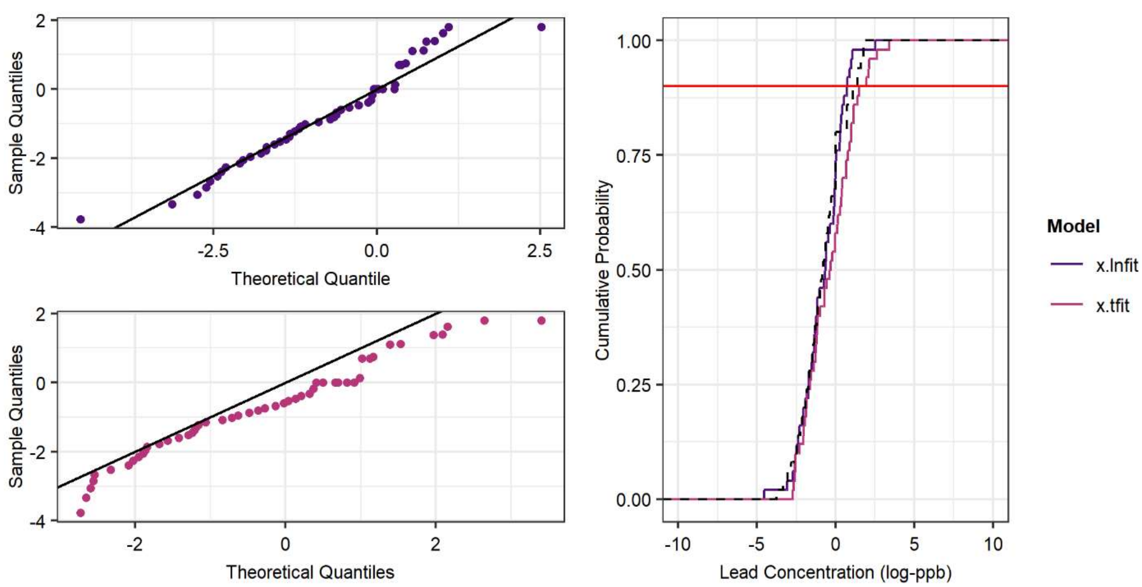

Figure SI.12. Utility B – 2004 Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

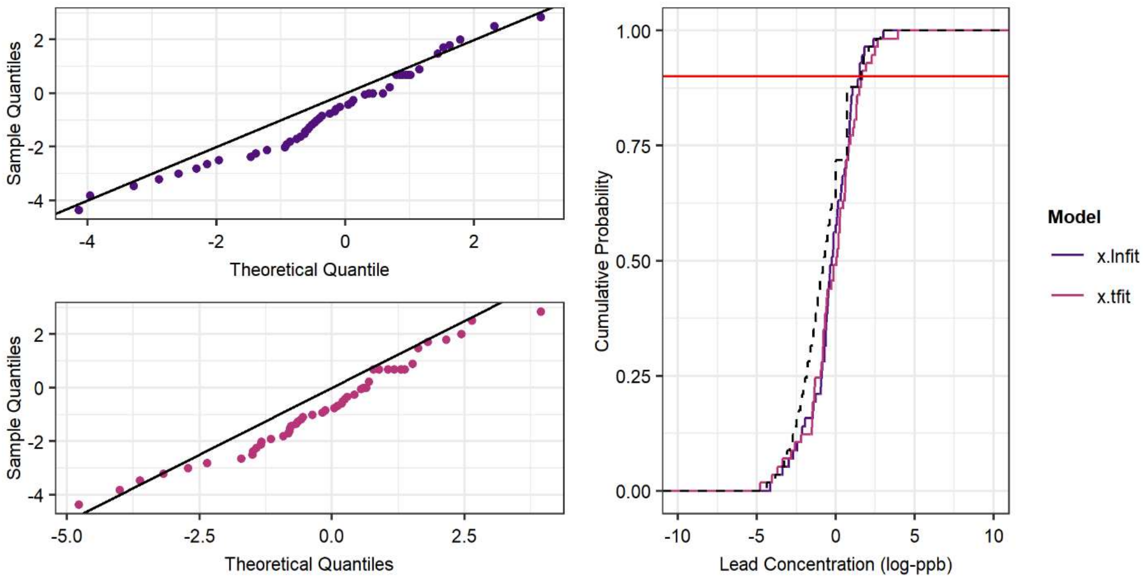

Figure SI.13. Utility B 2007 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

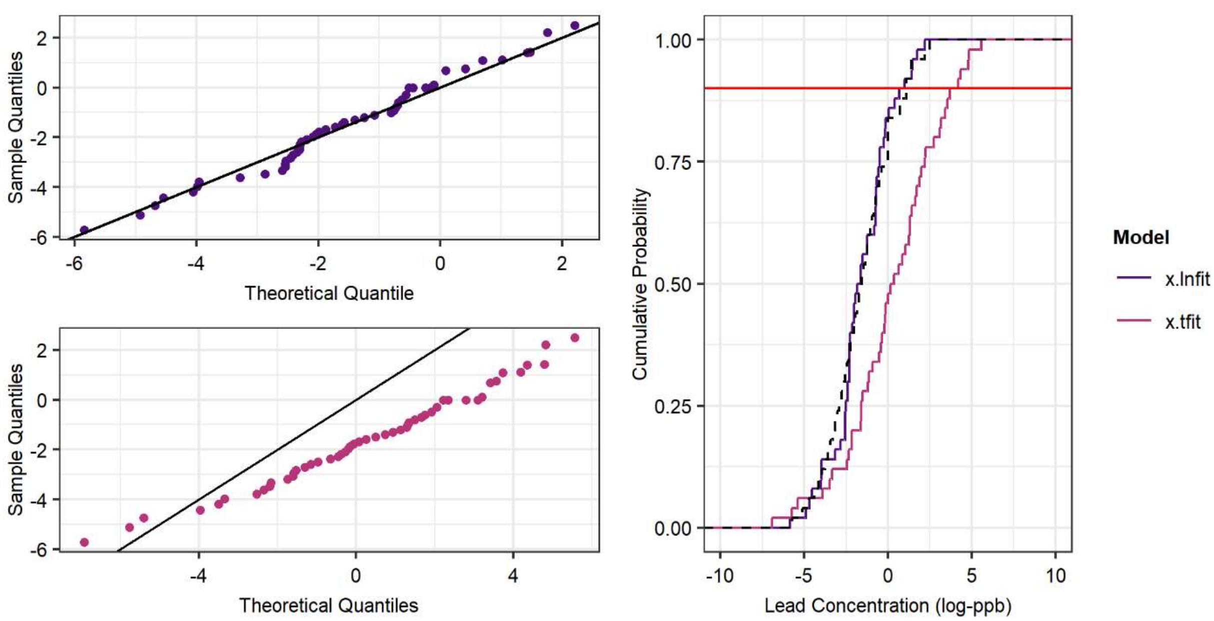

Figure SI.14. Utility B 2010 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

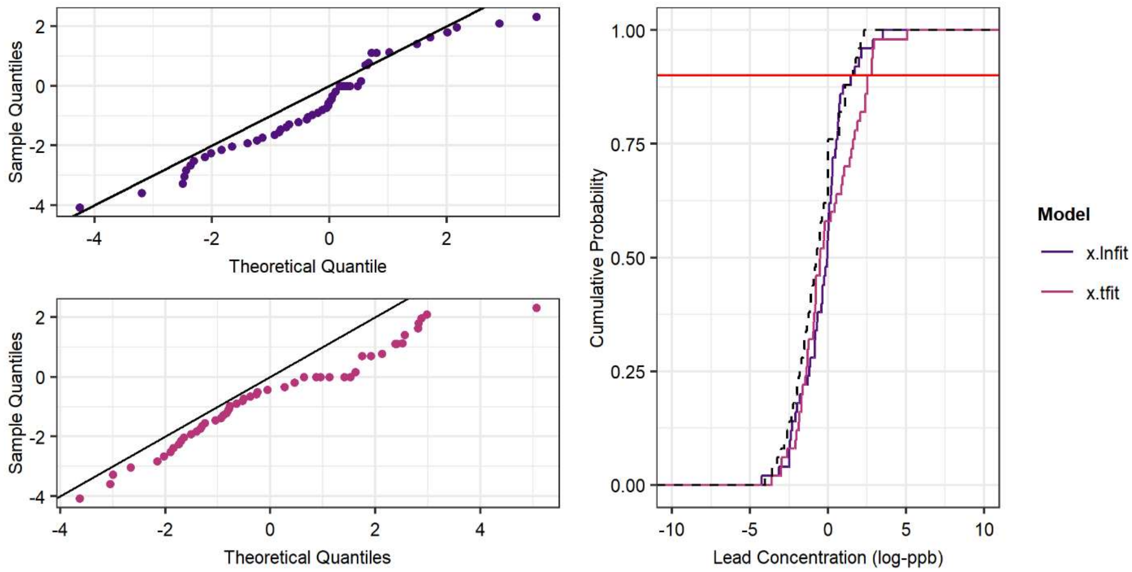

Figure SI.15. Utility B 2013 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

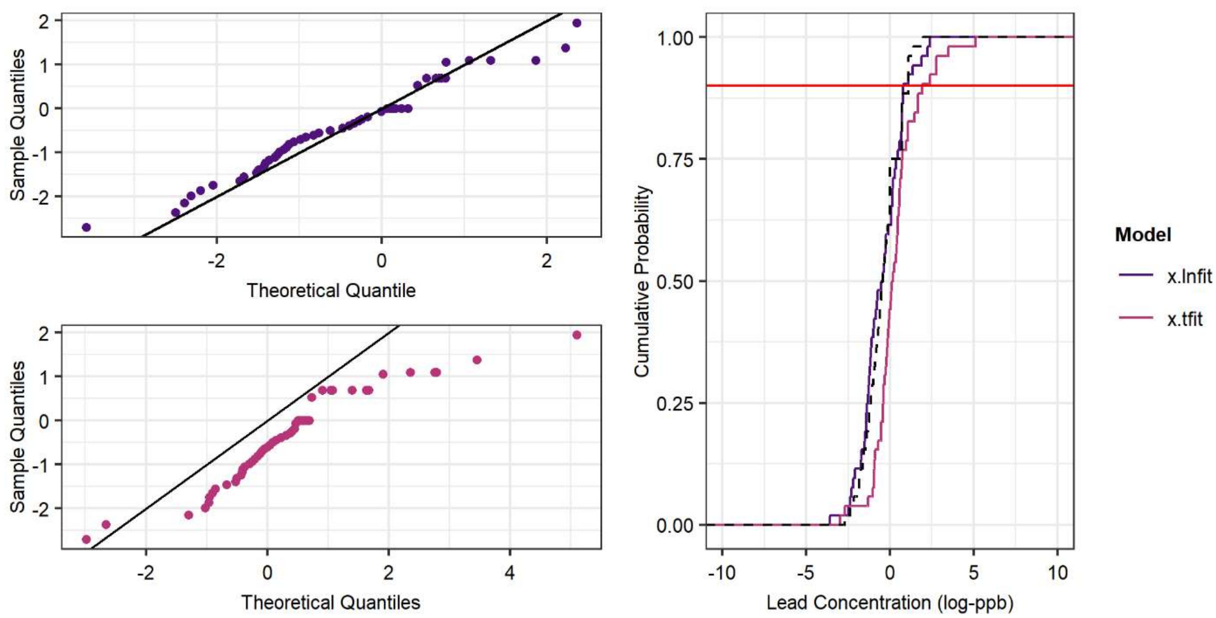

Figure SI.16. Utility B 2016 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

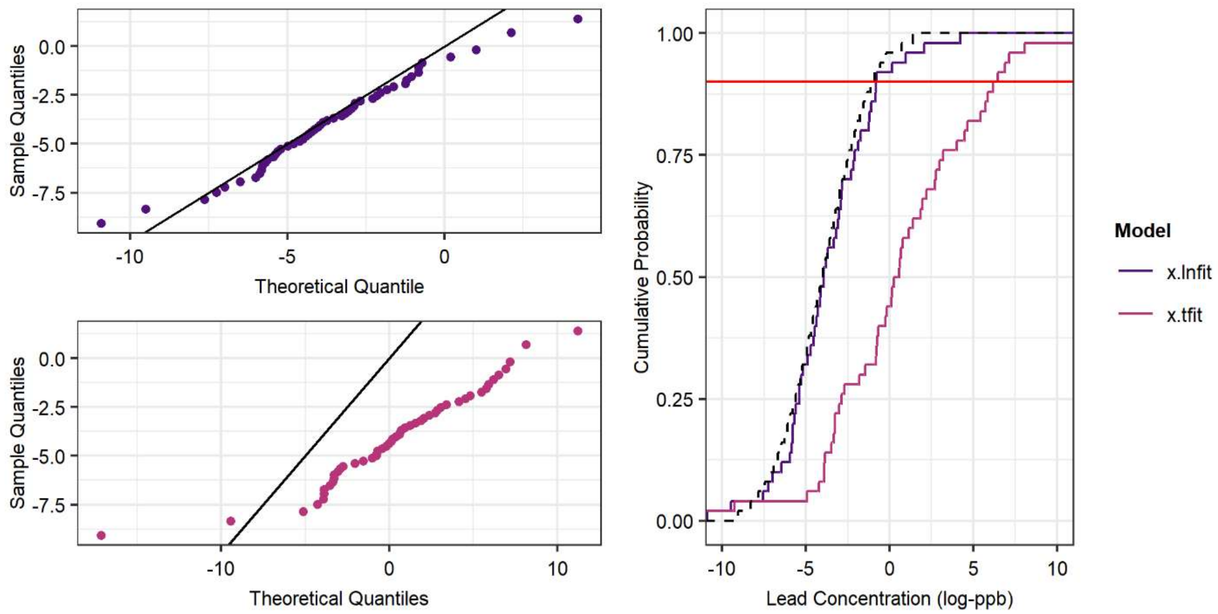

Figure SI.17. Utilityt C 2004 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

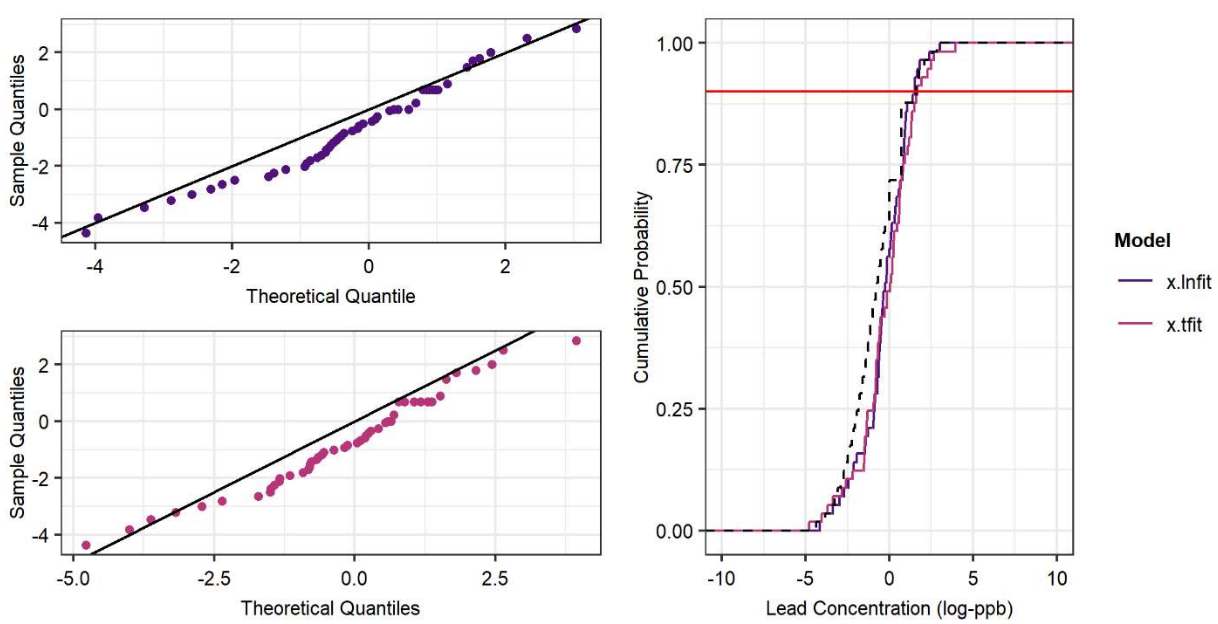

Figure SI.18. Utility C 2007 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

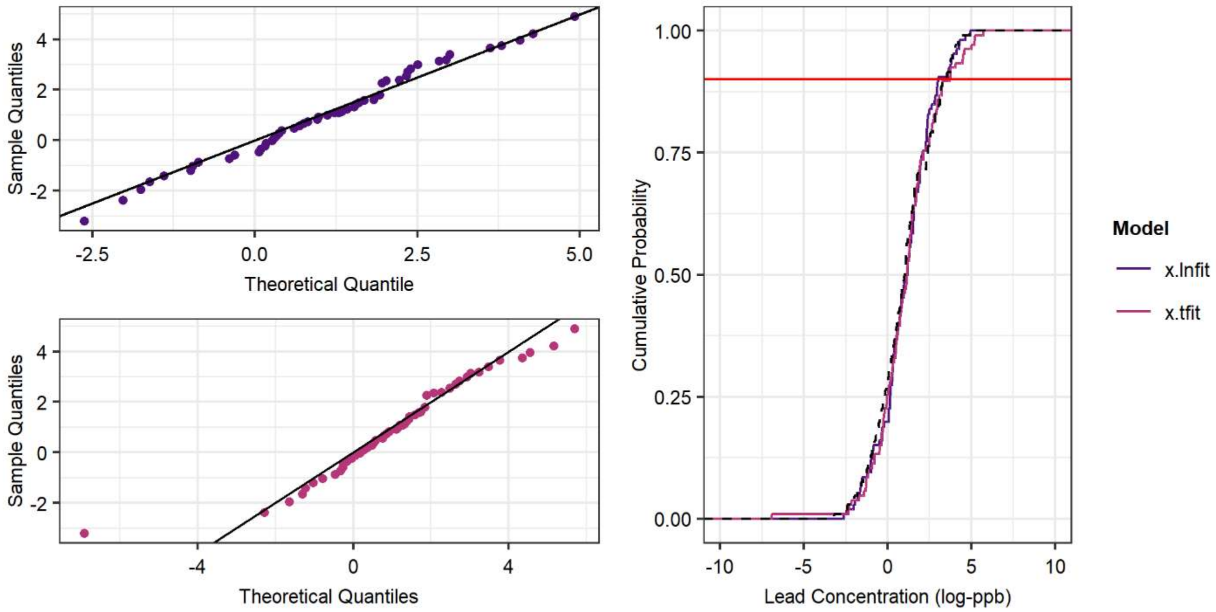

Figure SI.19. Utility C 2010 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

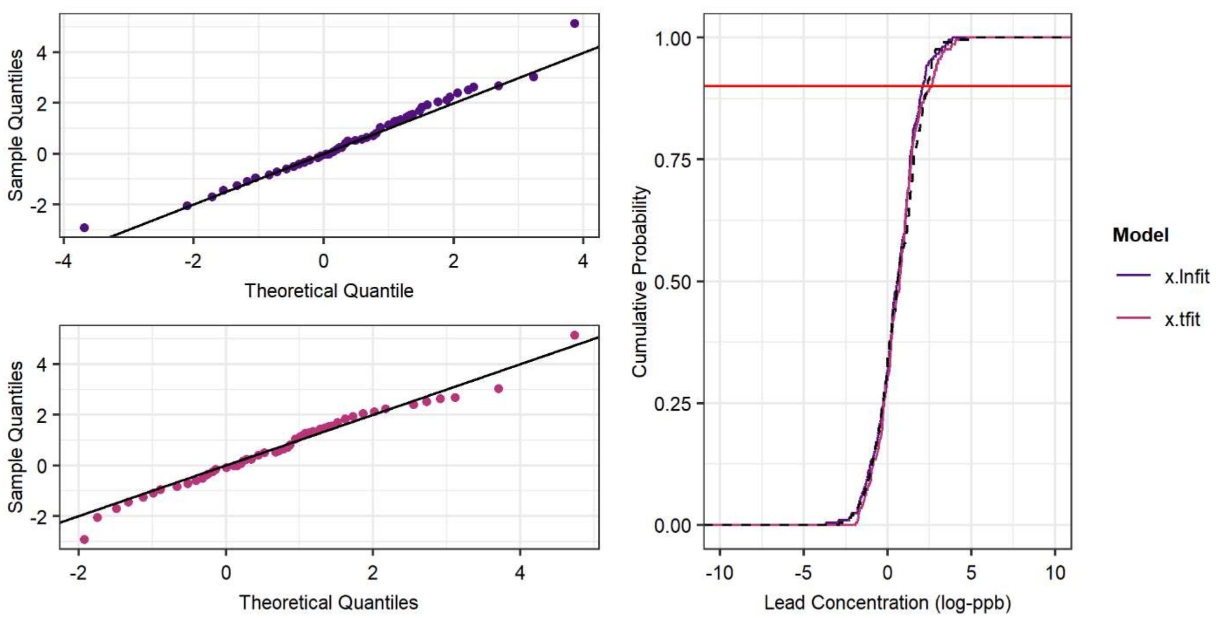

Figure SI.20. Utility C 2011 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

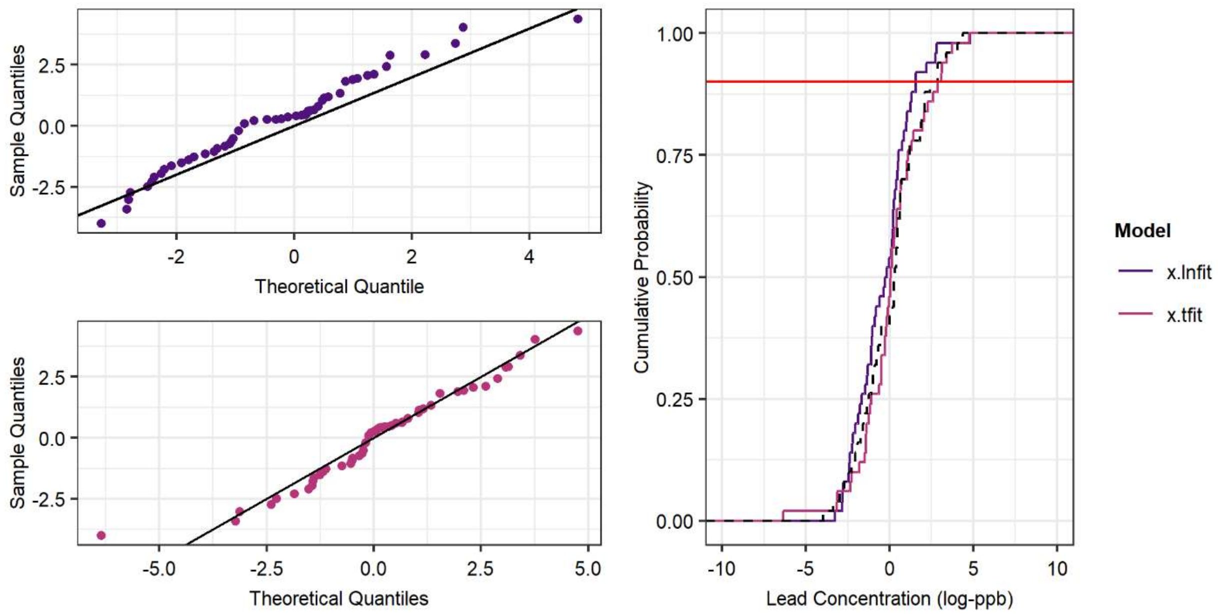

Figure SI.21. Utility C 2012 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

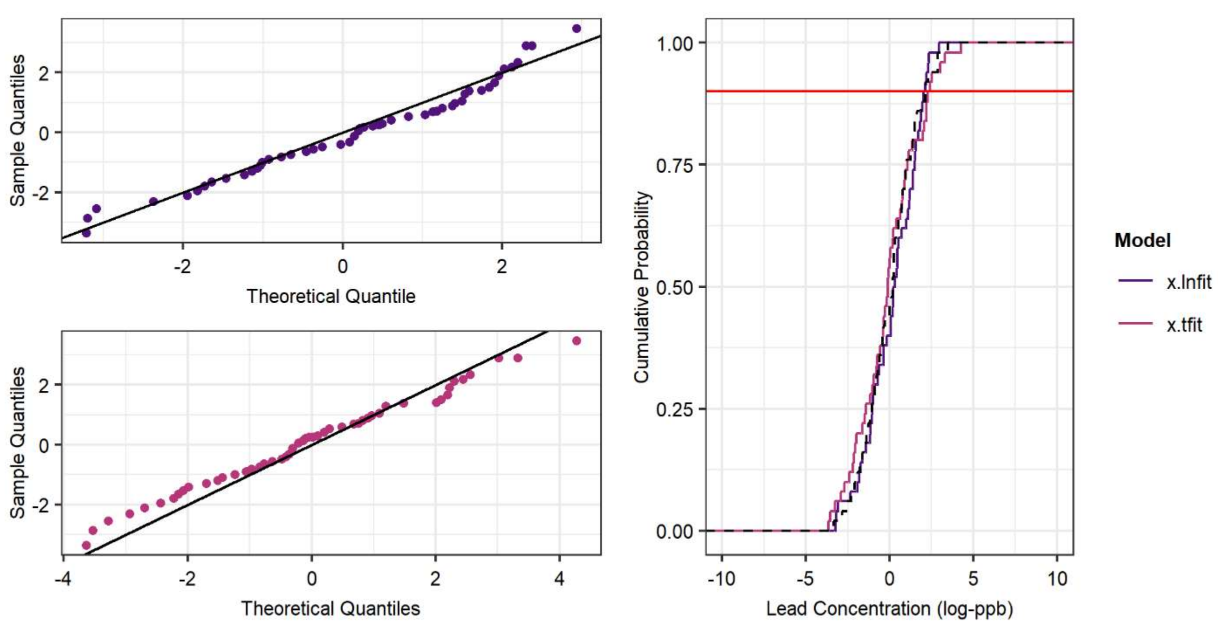

Figure SI.22. Utility C 2013 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

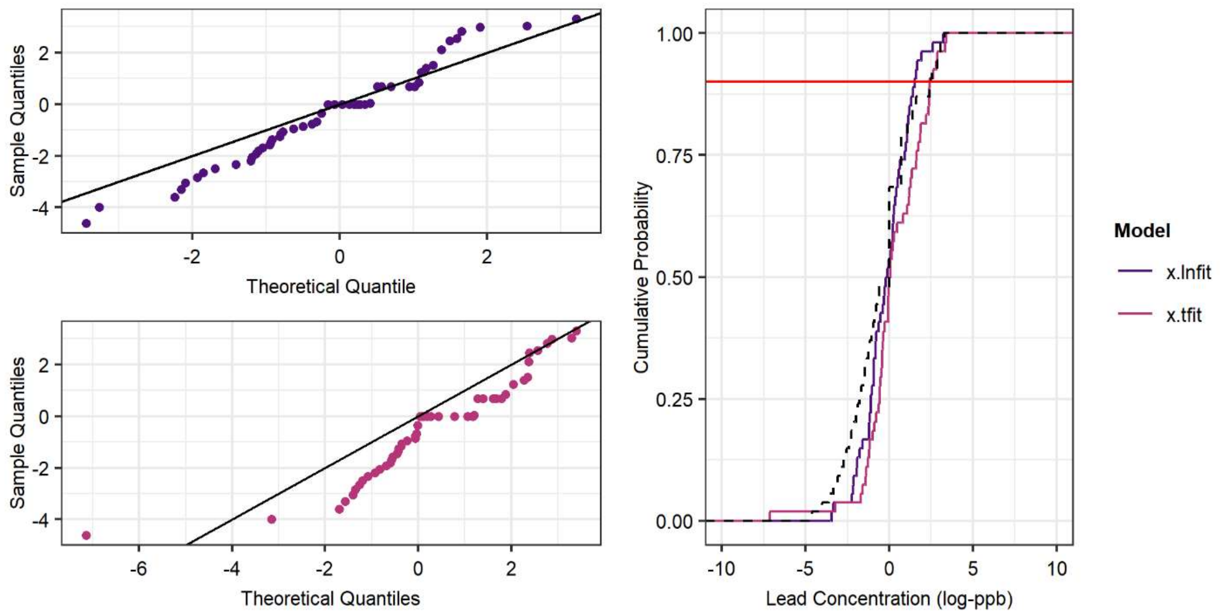

Figure SI.23. Utility C 2016 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

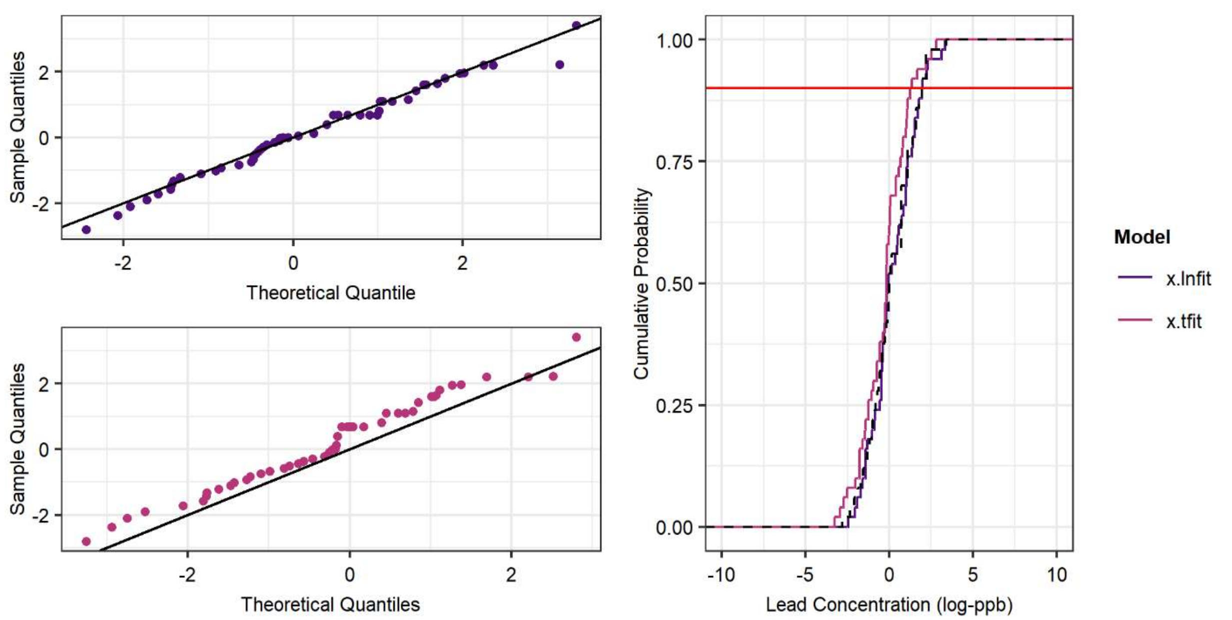

Figure SI.24. Utility D 2004 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

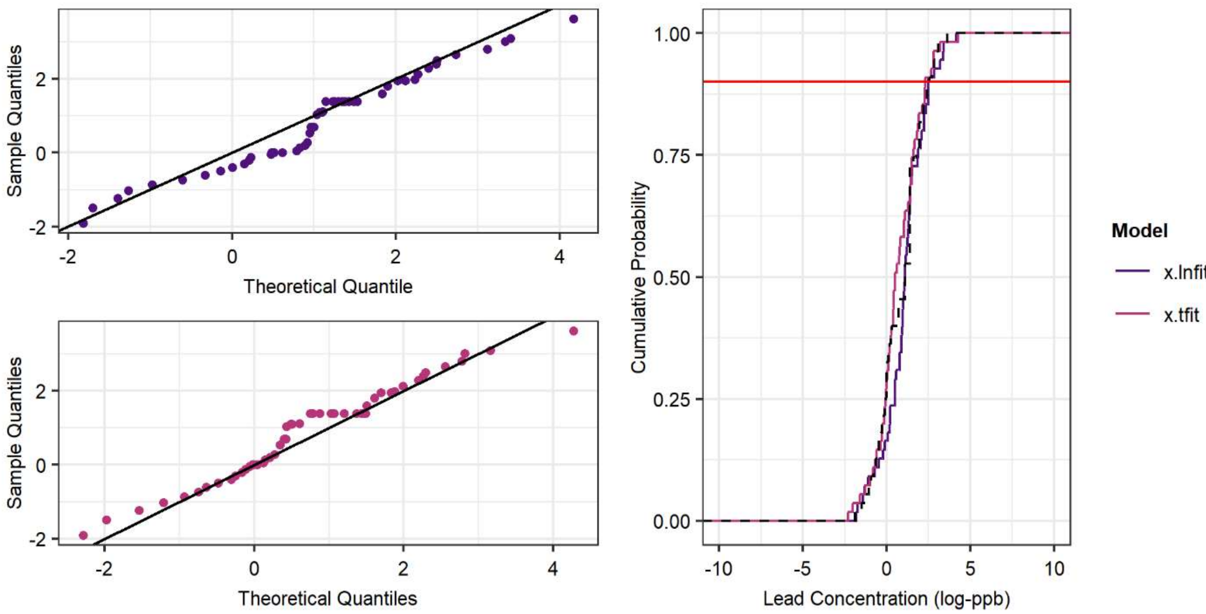

Figure SI.25. Utility D 2007 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

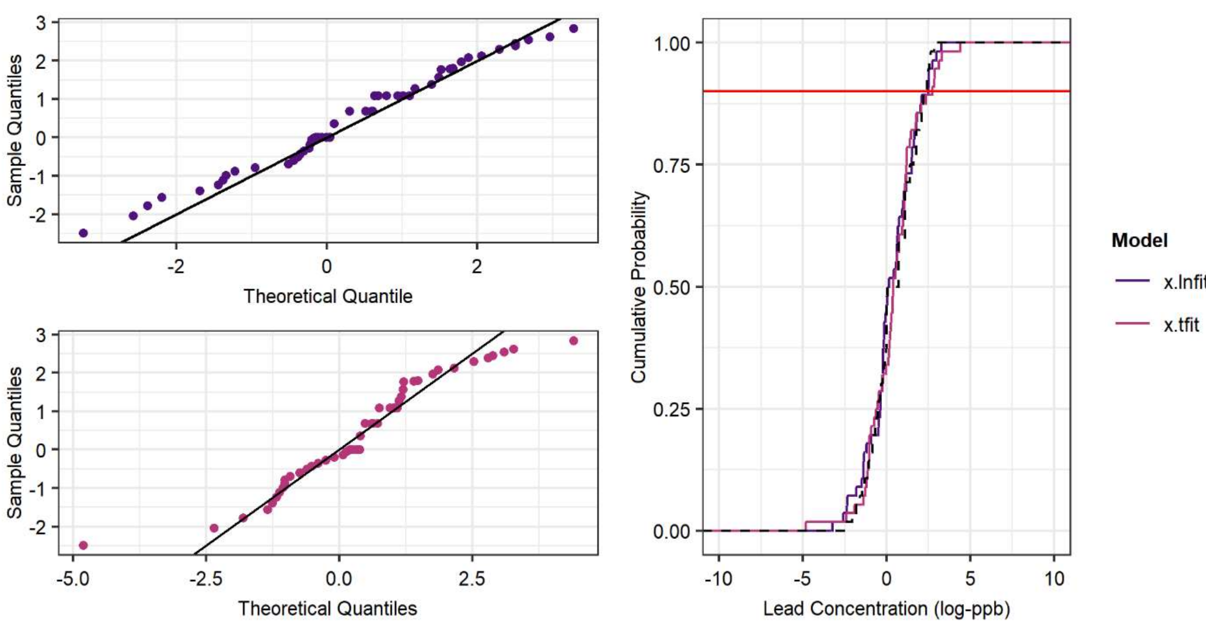

Figure SI.26. Utility D 2010 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

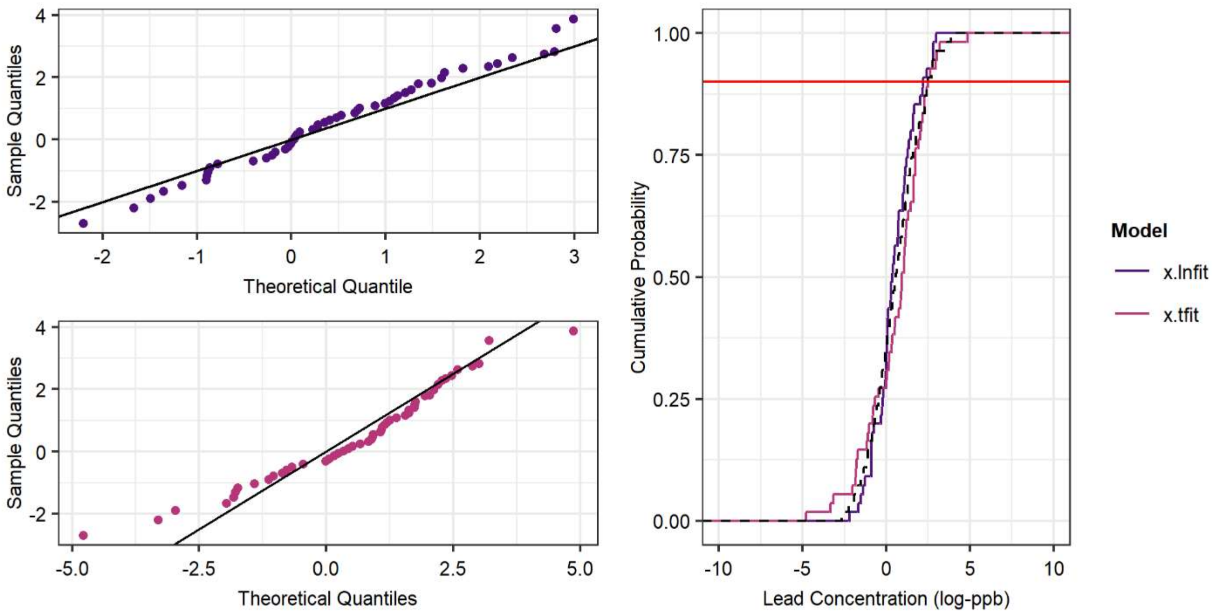

Figure SI.27. Utility D 2013 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

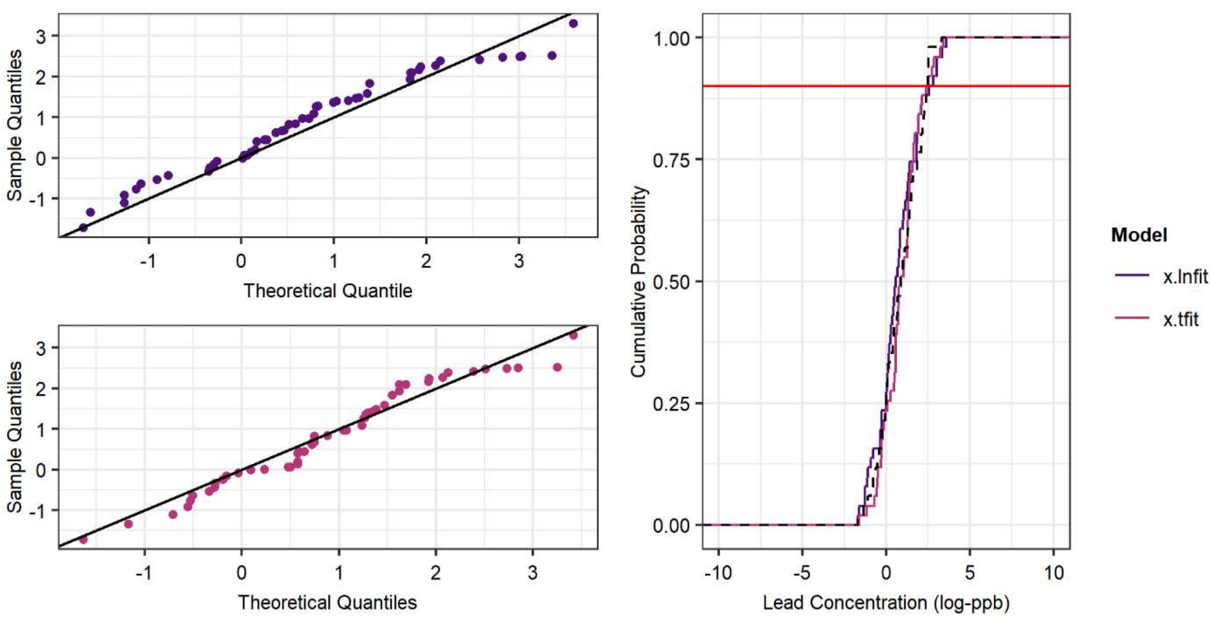

Figure SI.28. Utility D 2016 – Quantile-quantile plot of the normal and student's t-distribution compared to the interpolated dataset (left) and CDFs of the fitted distributions compared to the empirical distribution (right).

In addition, to the distribution fits evaluated through comparison of quantile-quantile plots and CDFs, four error metrics were compared and the root mean square error (RMSE) and the Bayesian Information Criterion (BIC) were selected. The mathematical formula for the BIC and the commonly u is shown in the following equation:

$$RMSE = \frac{\sum_{i=1}^{N} \sqrt{\left(\hat{y}_i - y_i\right)^2}}{N}$$

$$\begin{aligned}
&BIC = \ln\left(N\right)k - 2\ln\left(\hat{L}\right) \textit{ where} \\
&k \textit{ is number of parameters} \\
&N \textit{ is the number of samples} \\
&\hat{L} \textit{ is the loglikelihood}
\end{aligned}$$

Figure SI. 29 shows the results of the quantitative error analysis for each utility in each year for each model. While the RMSE showed the student's T-distribution model as having lower error values, the BIC metric, which is typically regarded as more robust, found the models to be tied, or the lognormal model to have a slightly lower error metric. Based on the BIC error metric and the visual inspection, the lognormal model, commonly used for environmental datasets, is identified as acceptable, and is used for all subsequent analyses.

Figure SI.30 shows the convergence of the simulation as the number of iterations increases. For each year and each utility the cumulative mean in the 90th percentile statistical is calculated as the number of bootstrap samples, *m,* increases. The mean stabilizes quickly, and by 200 iterations little variation is observed. 1000 bootstrap samples is used in these analyses and based on this analysis is sufficient.

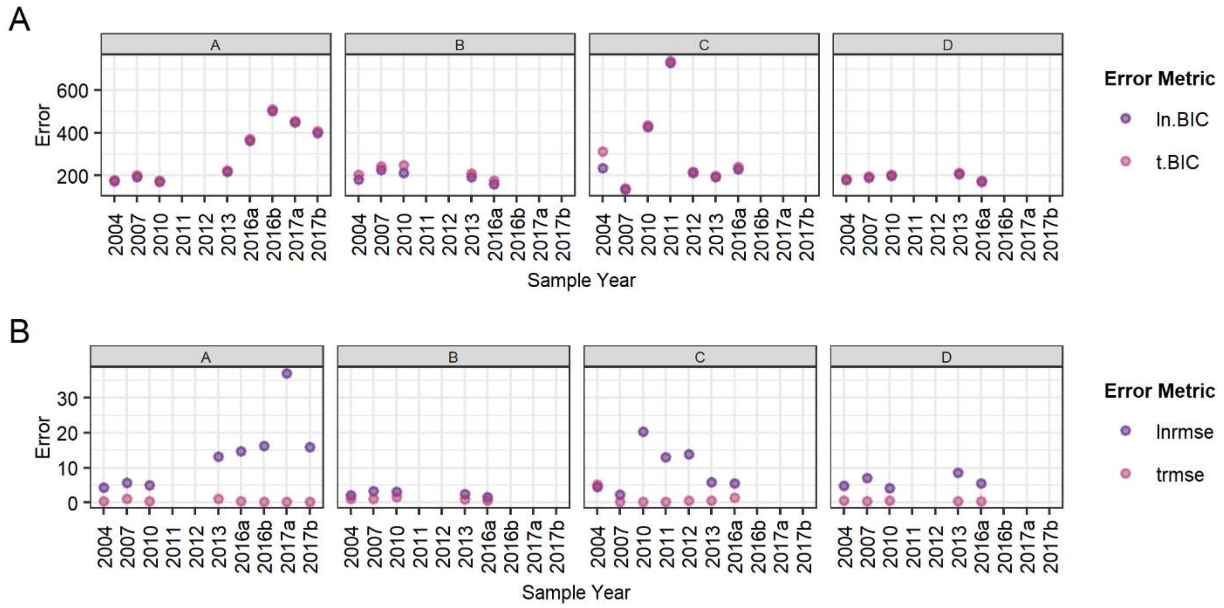

Figure SI.29. BIC and Root mean square error for utilities A through D from 2004 through 2017. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year.

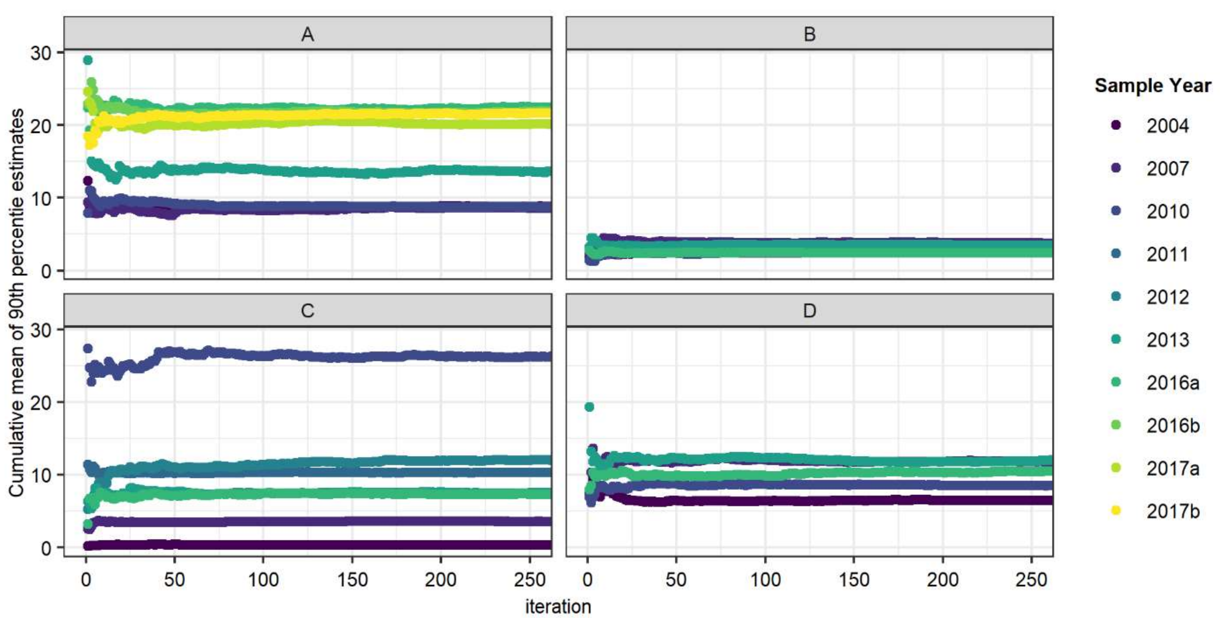

Figure SI.30: Variation of the mean of the 90th percentile distribution for utilities A through D from 2004 through 2017. Years designated as "a" or "b" indicate two sample periods were conducted that year. "a" represents samples collected between January and June and "b" represents samples collected between July and December of the indicated year. Convergence is determined at the number of iterations at which the cumulative mean of the distribution does not change as additional simulations are added.

4 Uncertainty assessment

The uncertainty in the 90th percentile shown in Figure 3 is based on parametric bootstrapping from the fitted model. This measure of uncertainty in the 90th percentile is compared to the uncertainty computed through sample bootstrapping. Figure SI.31 shows the CDFs for the 90th percentile calculated from a nonparametric bootstrap of the existing imputed datasets. Comparing these distributions provides a measure of the value associated with fitting a model and the potential drawbacks associated with not knowing any sample values beyond those in the LCR sample set. The dots in each figure show the reported LCR 90th percentile value for the original sample and where that value falls within the estimated distribution of the 90th percentile. The red line in each figure indicates the LCR AL.

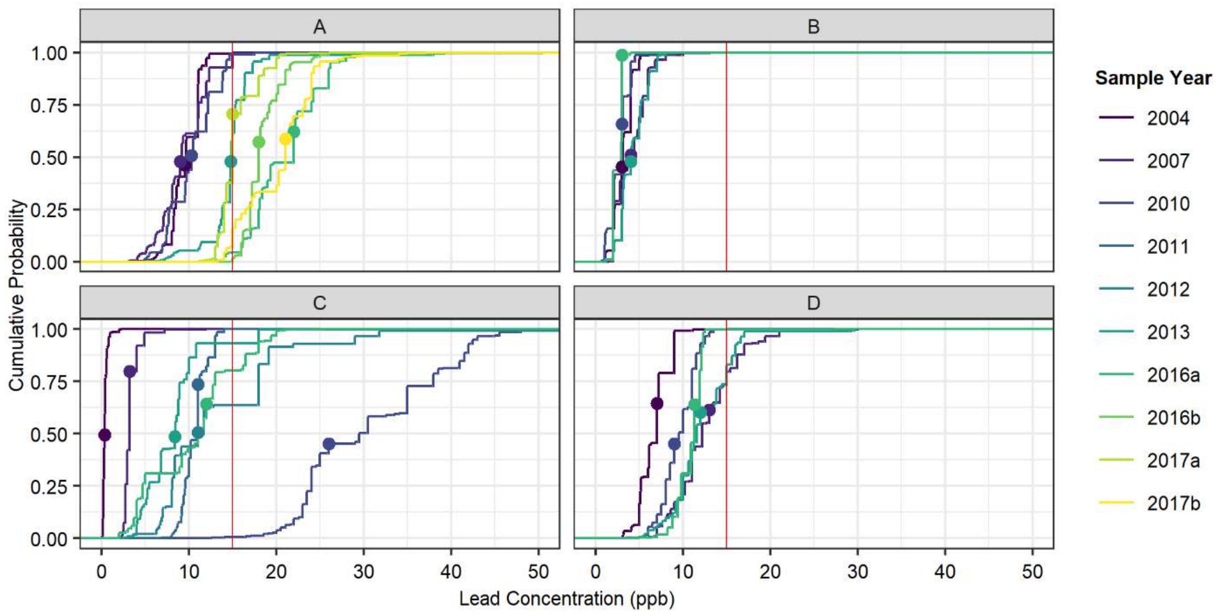

Figure SI.31. CDFs of uncertainty in imputed sample bootstrap for the 90th percentile for utilities A through D from 2004 through 2017. Years designated as "a" or "b" indicate two sample periods were conducted that year. "a" represents samples collected between January and June and "b" represents samples collected between July and December of the indicated year. Dots represent the LCR reported 90th percentile. Vertical red line is 15 ppb (the LCR AL).

Discontinuities in the distribution of the 90th percentiles predicted from the sample bootstrap approach are a result of restricting the predicted sampling results to values in the original sample (e.g., if no sample exists at 10 ppb in the original sample set then this sample cannot appear in the predicted distribution, even if it is a feasible result). Bootstrapping over such a small sample may cause a mischaracterization of portions of the distribution, where data are sparse, whereas the continuous fitted distribution is able to fill missing data gaps and provide a more accurate estimate of the 90th percentile value. Sampling from the fitted model also provides a more complete representation of the total population of lead results across all high-risk sites due to its continuous distribution of concentration results.

While the 90th percentile provides information about the extreme of the distribution, and is relevant for regulatory compliance, the median may provide a better approximation of the most-likely concentration associated with lead exposure for the high-risk sites sampled (by LCR requirements all have to include either a lead service line or lead solder). Figure SI.32 and Figure SI.33 show the uncertainty in the median, by parametric bootstrap and sample bootstrap methods, respectively. A large uncertainty range in the 90th percentile is not always indicative of a relatively wider range in uncertainty in the median. Comparison of the uncertainty in the median to uncertainty in the 90th percentile value indicates if there is an increase in the spread of the entire distribution versus just an elongation of the right tail of the distribution. For example, the years 2016 through 2017 for Utility A show similar increases in the uncertainty of both the 90th percentile value and the median, 2010 for Utility C in contrast only shows an increase in the 90th percentile uncertainty, but not the median indicating different changes in the lead concentration distribution leading to an exceedance of the LCR AL, and potentially a different underlying cause.

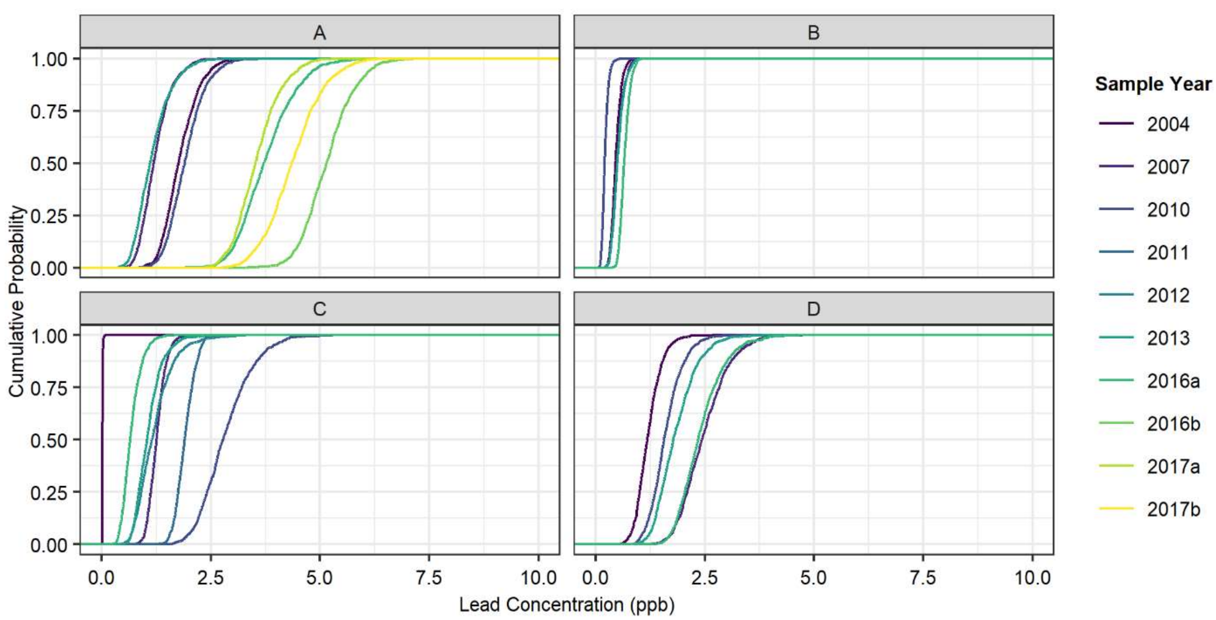

Figure SI.32. CDFs of the uncertainty in the median for utilities A through D from 2004 through 2017. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year. Estimated based on a parametric bootstrap of the fitted lognormal model.

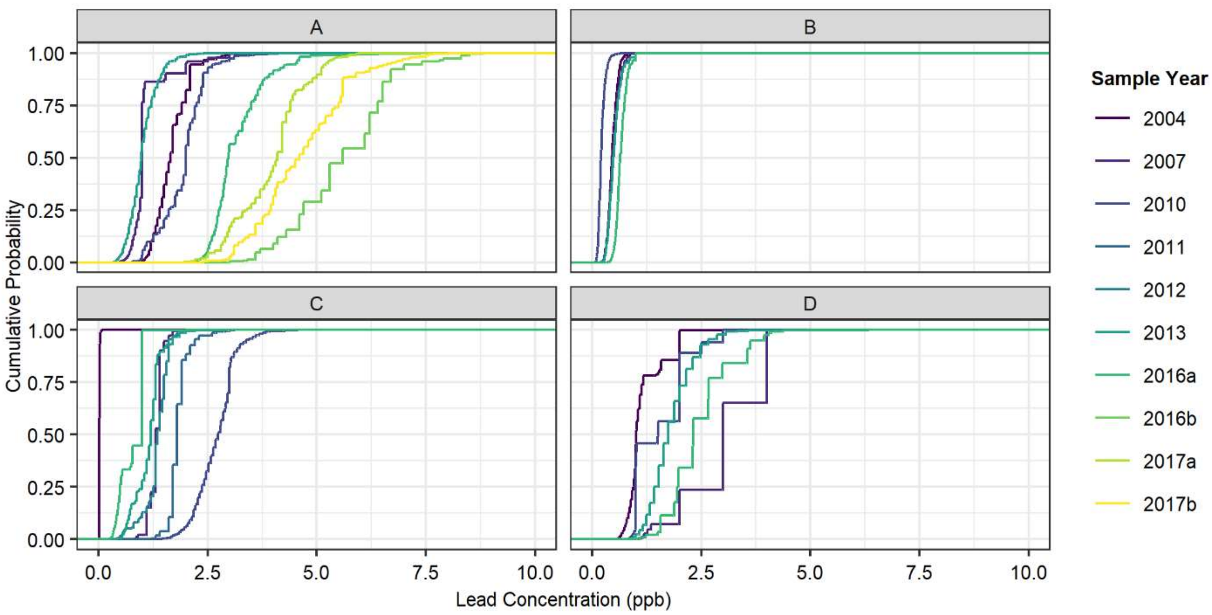

Figure SI.33. CDFs of the uncertainty in the median for utilities A through D from 2004 through 2017. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year. Estimated based on a sample bootstrap of the imputed dataset.

Figure SI.34 compares predictions of the median between the lognormal parametric distribution model and the non-parametric sample bootstrap model. Similar to the predictions of the 90th percentile statistics (Figure 5), predictions of median are close to well aligned and show limited deviation. The root mean square error (RMSE) is calculated to compare the models' median predictions and ranges from 0 to 0.74 across all sites and years. Table SI.2 shows the RMSE values for models in each year at each site. With both of these models there seems to limited uncertainty associated with the distribution model form selected.

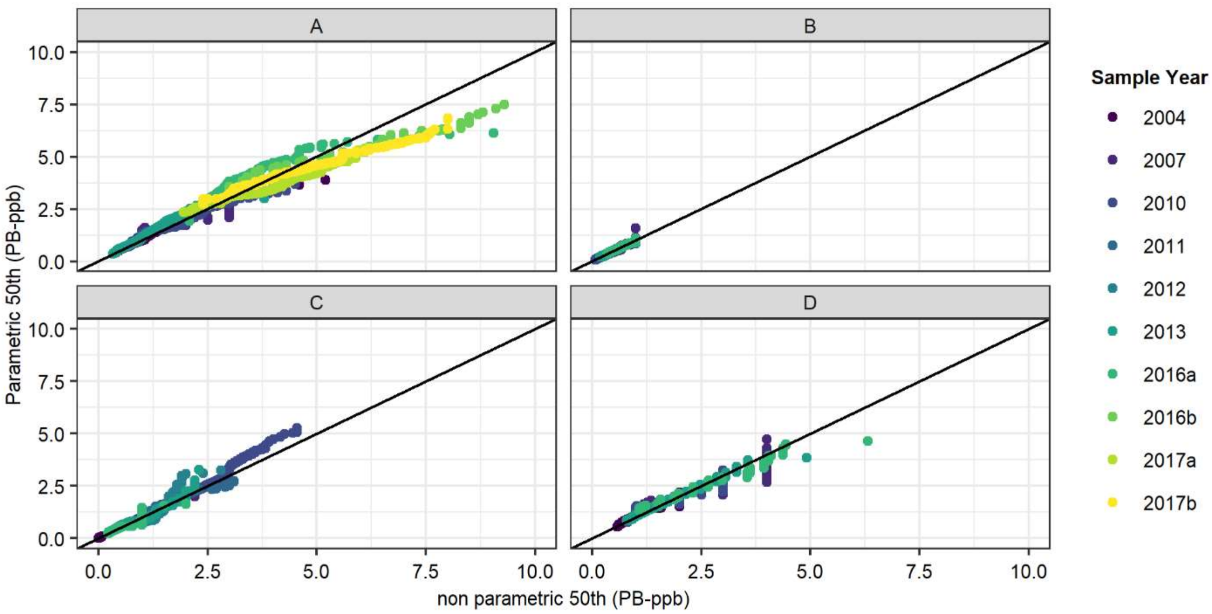

Figure SI.34. Comparison of median estimates between the parametric and non-parametric distribution models for utilities A through D from 2004 through 2017. Years designated as "a" or "b" indicate two sample periods were conducted that year. "a" represents samples collected between January and June and "b" represents samples collected between July and December of the indicated year.

Table SI.2. Root mean square errors calculated to compare model predictions of the 90th percentile and median.

Sample Year	Utility ID	RMSE: 90th percentile predictions	RMSE: Median predictions
2004	A	1.587	0.146

2007		1.676	0.271
2010		1.827	0.148
2013		4.303	0.175
2016a		2.188	0.628
2016b		3.754	0.698
2017a		4.746	0.501
2017b		1.860	0.513
2004		0.743	0.012
2007		0.690	0.031
2010	B	0.601	0.011
2013		0.948	0.020
2016		0.381	0.017
2004		0.100	0.001
2007		0.452	0.094
2010		5.303	0.230
2011	C	0.477	0.127
2012		2.722	0.259
2013		1.320	0.121
2016		4.045	0.189
2004		1.022	0.188
2007		0.853	0.738
2010	D	1.477	0.313
2013		1.320	0.069
2016		1.959	0.225

5 Difference tests and assessment of change over time

Many utilities may want to evaluate a change over time through comparison of sample sets collected over years. Plotting the trend in the 90th percentile is a tempting, but inaccurate way to evaluate a temporal trend. Statistical tests can be used to compare the distribution of datasets over time or compare changes in the median, which are more reliable ways to assess temporal change.

Each imputed dataset, made up of one sample year for one utility, is compared to all other years for that utility; two metrics were used to evaluate if temporal change occurred for each utility. First, the plotted empirical CDFs where compared visually (see Figure SI.35). Second, the non-parametric generalized Wilcoxon test[7] was used for a full distribution comparison. This test is selected as it compares the entirety of each distribution or dataset rather than a specific value such as the mean or median. The results of the Wilcoxon test allow for sample years to be sorted into groups and the results of the test compared to visual observations from comparison of the CDFs. A change in the median as measured by the results of the t-test are also compared to the empirical CDFs and any observed shift in the median. P-values are compared for the imputed samples directly in Table SI.3 and for the fitted lognormal models in Table SI.4. Cells shown in red have a p-value less than 0.05. This test shows no change among years for Utilities B and D, which is supportive of visual observations.

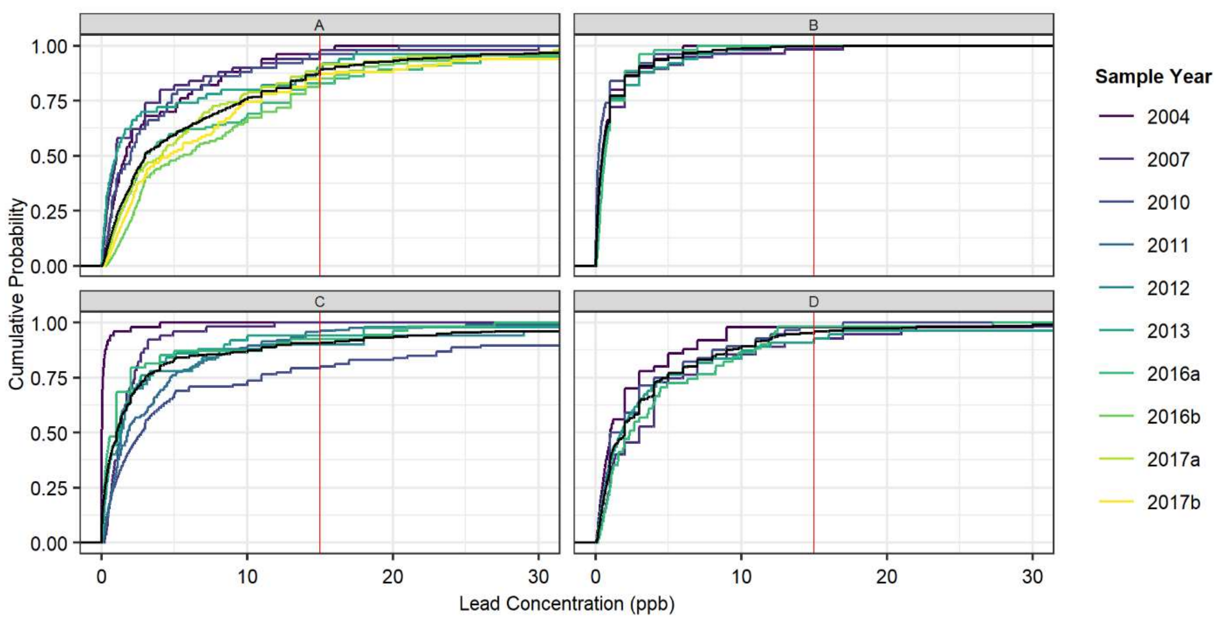

Figure SI.35. Empirical cumulative distribution function for utilities A through D from 2004 through 2017. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year.

Table SI.3. P-values from the nonparametric generalized Wilcoxon test comparing empirical distributions Utilities A through D from 2004 through 2017.

Utility ID		2007	2010	2011	2012	2013	2016a[1]	2016b[1]	2017a[1]	2017b[1]
A	2004	0.168	0.874			0.184	0.002	0.000	0.003	0.000
A	2007		0.126			0.907	0.000	0.000	0.000	0.000
A	2010					0.122	0.007	0.000	0.006	0.000
A	2013						0.000	0.000	0.000	0.000
A	2016a[1]							0.066	0.701	0.484
A	2016b[1]								0.008	0.216
A	2017a[1]									0.190
B	2004	0.896	0.025			0.788	0.158			
B	2007		0.024			0.915	0.236			
B	2010					0.020	0.001			
B	2013						0.306			
C	2004	0.000	0.000	0.000	0.000	0.000	0.000			
C	2007		0.002	0.023	0.749	0.478	0.019			
C	2010			0.067	0.006	0.001	0.000			
C	2011				0.049	0.016	0.000			
C	2012					0.860	0.140			
C	2013						0.136			
C	2014						0.001			
D	2004	0.007	0.250			0.146	0.009			
D	2007		0.090			0.239	0.995			
D	2010					0.701	0.108			
D	2013						0.268			

Note:

1. Years designated as “a” or “b” indicate two sample periods were conducted that year. “a” represents samples collected between January and June and “b” represents samples collected between July and December of the indicated year.

Table SI.4. P-values from the generalized Wilcoxon test comparing changes in the fitted distribution over time using the lognormal fit determined in the model fitting stage.

Utility ID		2007	2010	2011	2012	2013	2016a[1]	2016b[1]	2017a[1]	2017b[1]
A	2004	0.027	0.801			0.083	0.053	0.000	0.013	0.000
A	2007		0.096			0.723	0.000	0.000	0.000	0.000
A	2010					0.078	0.037	0.000	0.007	0.000
A	2013						0.000	0.000	0.000	0.000
A	2016a[1]							0.000	0.493	0.012
A	2016b[1]								0.001	0.264
A	2017a[1]									0.060
B	2004	0.370	0.169			0.528	0.212			
B	2007		0.022			0.163	0.727			
B	2010					0.454	0.010			
B	2013						0.060			
C	2004	0.000	0.000	0.000	0.000	0.000	0.000			
C	2007		0.045	0.313	0.131	0.003	0.000			
C	2010			0.145	0.004	0.000	0.000			
C	2011				0.015	0.000	0.000			
C	2012					0.506	0.012			
C	2013						0.038			
D	2004	0.001	0.198			0.417	0.018			
D	2007		0.017			0.008	0.153			
D	2010					0.752	0.216			
D	2013						0.151			

Note:

1. Years designated as "a" or "b" indicate two sample periods were conducted that year. "a" represents samples collected between January and June and "b" represents samples collected between July and December of the indicated year.

6 References

(1) US EPA. *Lead and Copper Rule*; 1991; Vol. 40 Part 141.1.

(2) US EPA. *Proposed Lead and Copper Rule Revisions*; 2019; Vol. 40 CFR Parts 141 and 142, pp 61684–61774.

(3) Del Toral, M. A.; Porter, A.; Schock, M. R. Detection and Evaluation of Elevated Lead Release from Service Lines: A Field Study. *Environmental Science & Technology* **2013**, *47* (16), 9300–9307. https://doi.org/10.1021/es4003636.

(4) Trueman, B. F.; Camara, E.; Gagnon, G. A. Evaluating the Effects of Full and Partial Lead Service Line Replacement on Lead Levels in Drinking Water. *Environmental Science & Technology* **2016**, *50* (14), 7389–7396. https://doi.org/10.1021/acs.est.6b01912.

(5) Deshommes, E.; Bannier, A.; Laroche, L.; Nour, S.; Prevost, M. Monitoring-Based Framework to Detect and Manage Lead Water Service Lines. *Journal - American Water Works Association* **2016**, *108*, E555–E570. https://doi.org/10.5942/jawwa.2016.108.0167.

(6) Pennsylvania Department of Environmental Protection. Drinking Water Reporting System http://www.drinkingwater.state.pa.us/dwrs/HTM/Welcome.html (accessed Sep 7, 2018).

(7) Helsel, D. R. *Statistics for Censored Environmental Data Using Minitab and R*, 2nd ed.; Wiley series in statistics in practice; Wiley: Hoboken, N.J, 2012.

(8) Huston, C.; Jurarez-Colunga, E. *Guidelines for Computing Summary Statistics for Data-Sets Containing Non-Detects*; Simon Fraser University: British Columbia, 2009.

(9) Titterington, D. M.; Smith, A. F.; Makov, U. E. *Statistical Analysis of Finite Mixture Distributions*; Probability and mathematical statistics: Applied probability and statistics; Wiley, 1985; Vol. 198.

(10) Baron, M. *Probability and Statistics for Computer Scientists*, 2nd ed.; CRC Press: Boca Raton, FL, 2014.

(11) Burlingame, G. A.; Sandvig, A. How to Mine Your Lead and Copper Data. *Opflow* **2004**, *30* (6), 16–19. https://doi.org/10.1002/j.1551-8701.2004.tb01760.x.