Environmental Science & Technology Letters

Letter

pubs.acs.org/journal/estlcu

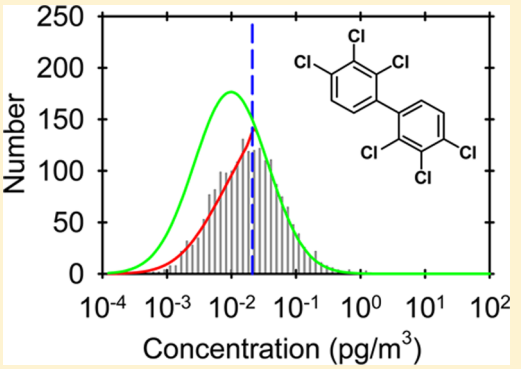
A Statistical Approach for Left-Censored Data: Distributions of Atmospheric Polychlorinated Biphenyl Concentrations near the Great Lakes as a Case Study

Ronald A. Hites*

School of Public and Environmental Affairs, Indiana University, Bloomington, Indiana 47405, United States

Supporting Information

ABSTRACT: Polychlorinated biphenyl (PCB) congener concentrations were measured in atmospheric samples collected once every 12 days at six sites on the shores of the North American Great Lakes. These data have been obtained as part of the Integrated Atmospheric Deposition Network (IADN), which began in 1991. This data set now consists of ~2900 samples, each of which has been measured for ~80 PCB congeners. Some of these congeners are present at levels sufficiently high to be quantitated in almost every sample, while others are detected in fewer than two-thirds of the samples. These latter congeners represent an example of left-censored environmental measurements. This paper offers a simple approach to dealing with uncensored and censored atmospheric PCB concentration data based on a careful examination of the distribution function of the data and using the curve fitting power of the Solver feature of Excel.



INTRODUCTION

Imagine that one has 30 measurements of the atmospheric concentrations of several polychlorinated biphenyl (PCB) congeners taken every year over a 20 year period, giving a total of 600 measurements of each congener. Further, imagine that one wants to compare these measurements as a function of time to determine if the atmospheric levels are, presumably, decreasing as a function of time and, if so, how fast. One way to do this is to calculate the annual average concentration of each congener and to regress these averages against time. To calculate that average, it is important to understand the distribution function of the data. Only when one has this knowledge can one use the right measure of central tendency and variability. While it is commonly assumed that environmental measurements are log-normally distributed,[1] there are few detailed examinations of this assumption in the literature. There is an additional complication. Some PCB data sets, particularly those for air sampled at remote locations, are populated with either empty cells or values listed as less than some limit of detection. Thus, one might find data sets in which some congener has valid measurements for, say, two-thirds of the samples and has missing measurements for the others. This effect is known as left censoring, and the problem is to find the correct central tendency given that, in this case, one-third of the measurements are not known.

One approach to dealing with censored data is to replace the empty cells with finite values. Such replacements have included inserting the detection limit itself, half of this limit, or for reasons that are not clear, this limit divided by the square root of two.[2] None of these simple replacements are intellectually satisfying and, if followed routinely, would lead to artificially inflated average concentrations. There are other ways of dealing with censored data: there is even a book on the subject.[3] However, these approaches are usually complex and are not widely used. This paper offers a simple approach to dealing with uncensored and censored atmospheric PCB concentrations based on a careful examination of the distribution function of the data.

This paper focuses on PCB concentrations measured in atmospheric samples collected once every 12 days at six sites on the shores of the North American Great Lakes. These data have been obtained as part of the Integrated Atmospheric Deposition Network (IADN), which started its operations in 1991. This data set now consists of ~2900 samples, each of which has been measured for ~80 PCB congeners. Some of these congeners are present at levels sufficiently high to be quantitated in almost every sample (congeners 18, 52, and 101, for example), while others (congeners 100, 128, and 201, for example) are detected in fewer than two-thirds of the samples. These latter congeners represent an example of left-censored environmental measurements caused by the analytical method not being sensitive enough to determine accurate concentrations for PCB congeners at the lowest levels.

Received: July 21, 2015
Revised: August 25, 2015
Accepted: August 26, 2015

DOI: 10.1021/acs.estlett.5b00223
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

Table 1. Summary of the Geometric Means from the Descriptive Statistics, from the Fitted Normal Distributions Shown in Figure 1 Using eq 2, from the Uncensored Fitted Distributions Shown in Figure 2 Using eq 5, and from the Uncensored Fitted Distributions Shown in Figure 3 Using eq 5 and Using 20% Randomly Selected Data[a]

PCB congener	% detected	geo mean from descriptive statistics	median from descriptive statistics	geo mean from eq 2 (see Figure 1)	geo mean from eq 5 (see Figure 2)	geo mean from eq 5 using 20% of the data (see Figure 3)	LOQ from eq 5 (see Figure 2)	LOW using 20% of the data (see Figure 3)
PCB-18	99.8	1.31	1.27	1.26	1.25	1.16	0.26	0.26
PCB-52	100.0	0.846	0.835	0.821	0.817	0.792	0.150	0.150
PCB-101	100.0	0.467	0.465	0.453	0.448	0.468	0.093	0.100
PCB-19	84.1	0.0943	0.0951	0.0944	0.0732	0.0695	0.0608	0.1003
PCB-83	79.6	0.0269	0.0268	0.0267	0.0192	0.0193	0.0187	0.0161
PCB-180	87.5	0.0428	0.0415	0.0407	0.0362	0.0354	0.0151	0.0175
PCB-100	67.3	0.0494	0.0496	0.0486	0.0304	0.0302	0.0580	0.0420
PCB-128	65.0	0.0164	0.0164	0.0163	0.0089	0.0092	0.0209	0.0190
PCB-201	55.6	0.0226	0.0235	0.0239	0.0147	0.0141	0.0511	0.0413

[a]In the latter two cases, the fitted limits of quantitation (LOQ) from eq 5 are also given. All values are in picograms per cubic meter.

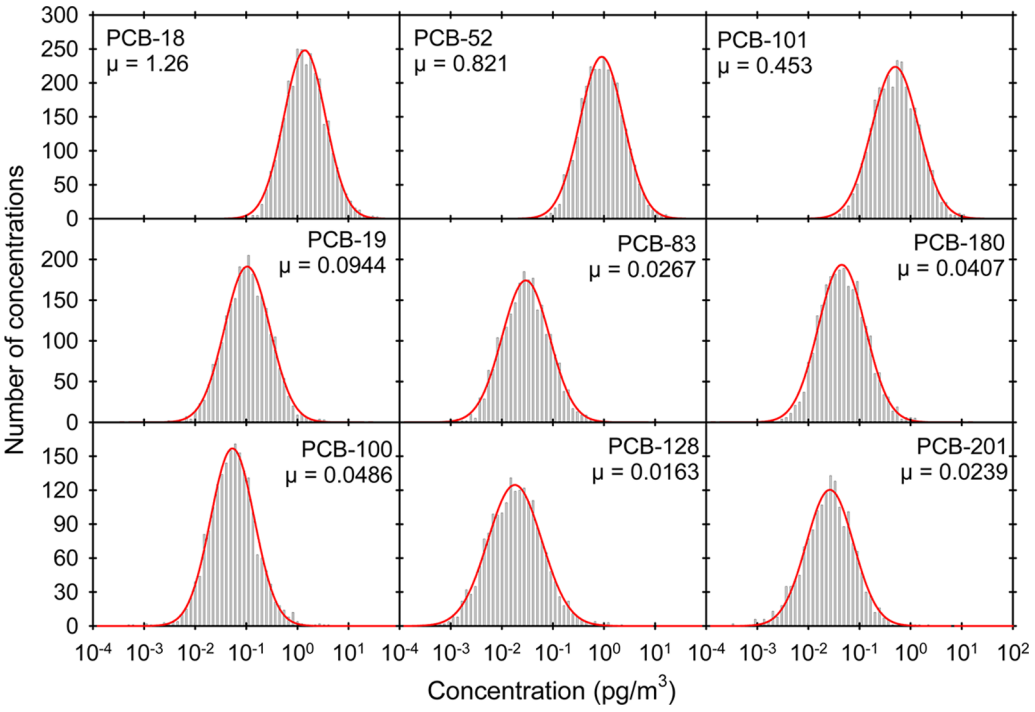


Figure 1. Histograms of the concentrations of nine PCB congeners measured in the atmosphere at six sites on the shores of the North American Great Lakes. The *y*-axis shows the number of measurements in a given bin, and the *x*-axis shows the atmospheric concentration on a logarithmic scale. The red line was fitted using eq 2. The geometric means (μ) are given for each congener. The parameters of these fits are listed in Table S2.

EXPERIMENTAL SECTION

The details of the sample collection and PCB analysis procedures have been published previously;[4] thus, only a summary is given here. The samples are collected at six sampling sites for 24 h once every 12 days (every 24 or 36 days at Point Petre). The temporal coverage of the data sets is different at the different sites, but in general, there are ~30 samples each year. The locations of the sampling sites and the dates of sampling are given elsewhere.[5] The air is sampled by a high-volume sampler at a flow rate such that ~820 m^3 is sampled over the 24 h period. The air is first pumped through a 2.2 μm filter to collect the particles and then through a bed of XAD-2 resin to collect the vapor phase components. Once returned to the laboratory, the particle and vapor phase media are extracted separately, and the extracts are cleaned up and analyzed separately. The PCBs are present in only the vapor phase, and PCBs in these samples are measured with electron capture gas chromatography with a 60 m long column. All analyses are based on internal calibration standards. Extensive quality assurance/quality control procedures have been implemented.[6]

To make this discussion tractable, I am only going to focus on PCB congeners 18, 52, and 101, which are detected in almost every sample; congeners 19, 83, and 180, which are detected in 80−88% of the samples; and congeners 100, 128, and 201, which are detected in 56−67% of the samples. In each case, the number of measurements is given in Table S1.

We have previously noted that these atmospheric PCB concentrations can be modeled by a harmonic regression of the form[7]

$$\ln(C_t) = a_0 + a_1 \sin(zt) + a_2 \cos(zt) + a_3 \log^2(\text{pop}) + a_4 t \quad (1)$$

DOI: 10.1021/acs.estlett.5b00223
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

where C_t is the PCB atmospheric concentration (in picograms per cubic meter) on date t, $z = 2\pi/365.25$ (which fixes the periodicity to 1 year), *pop* is the number of people living and working within a 25 km radius of the sampling site, and the a_i values are constants fitted using a multiple-regression analysis. To put all of the PCB concentrations on the same scale, I fitted each data set using eq 1 and subtracted the $a_3 \log^2(\text{pop})$ term from each $\ln(C_t)$ value. This removed the effect of population near the sampling sites and allowed me to pool all of the data from all of the sites. The fitted values of a_3 are listed in Table S1. The remaining variability is due to seasonal effects (which are large), temporal changes (which are small), and measurement error (which is approximately ±20%). After this correction, the effective analytical detection limit was ∼0.001 pg/m^3 for most congeners. This level is based on the lack of a gas chromatographic peak in the raw data of sufficient signal-to-noise to provide a valid measurement. Concentrations below this limit are empty cells in the spreadsheets and were not used in the following analysis.

RESULTS AND DISCUSSION

Because I suspected that the data were log-normally distributed, my first step in analyzing these PCB concentration data was to convert each measurement to its logarithm; in this case, I used the natural logarithm so that the constant a_4 in eq 1 represents a rate constant. Table S1 gives the descriptive statistics of all of these $\ln(C_t)$ values for each of the nine PCB congeners on which I have elected to focus. As noted above, I have divided them into three groups: those detected in 100% of the samples (congeners 18, 52, and 101), those detected in ∼84% of the samples (congeners 19, 83, and 180), and those detected in ∼63% of the samples (congeners 100, 128, and 201). The geometric means listed in Table S1 are the antilogarithms of the means of the $\ln(C_t)$ values, but the standard deviations are not. Columns 3 and 4 of Table 1 summarize the geometric means and the medians of these data. Note that the geometric means are similar to the medians of these distributions, as expected for log-normal distributions.

Next, the number of measurements in a specific $\ln(C_t)$ range was tabulated. For all congeners, the $\ln(C_t)$ values ranged from −9.0 to 9.0, and the bin size was 0.2, giving 90 bins. The histograms for the nine PCB congeners are plotted as a function of C_t on a logarithmic scale in Figure 1 (please ignore the red lines for the moment). For example, 250 measurements of PCB-18 were in the concentration range of 1.000−1.221 pg/m^3, which corresponds to a bin range of 0.000−0.200 on a logarithmic scale. The histograms for all nine congeners suggest that these data are normally distributed; thus, the histograms were fitted with a normal distribution using

$$m(x_i) = b_0 \exp[-(x_i - b_1)^2/2b_2^2] \quad (2)$$

where $x_i = \ln(C_t)$, $m(x_i)$ is the number of such measurements in each bin, b_0 is a scale factor, b_1 is the mean of the distribution, and b_2 is its standard deviation. The resulting fitted values are shown in Figure 1 as the red lines, and the fitted parameters are listed in Table S2. The geometric mean is given by

$$\mu = \exp(b_1 - 0.1) \quad (3)$$

The subtraction of 0.1 unit (half of the bin size) is necessary because x_i is a discrete variable, but the fitted function (eq 2) is a continuous function. These geometric means are summarized in column 5 of Table 1. As expected, the geometric means from this curve fitting are almost the same as those from the descriptive statistical analysis (compare columns 3 and 5 of Table 1). The small differences that remain are likely due to the discrete versus continuous variable correction discussed above.

It is clear from Figure 1 that all of these data are log-normally distributed. This is true even for those PCB congeners that are detected in fewer than two-thirds of the samples (see PCB-100, -128, and -201). This observation suggests that there is no fixed threshold below which the PCBs are not detected; rather, there is a gradual diminution in the detectability of these compounds as their concentrations decrease, reaching zero at the analytical detection limit of ∼0.001 pg/m^3.

In addition to the central tendency and the width of the distributions, two other measures of its shape are helpful. These are the distribution's skewness and kurtosis. Skewness is the third moment of the counts about the mean, and kurtosis is the fourth moment about the mean. For a perfectly normal distribution, both of these values would be zero. The skewness and kurtosis of the raw data are listed in Table S1 for each of the nine congeners. Skewness evaluates the symmetry of the distribution, and for all of the PCB congeners discussed here, the skewness is small. The kurtosis evaluates the "peakedness" or "pointiness" of the distribution; a relatively high, positive kurtosis indicates that the data in the middle of the distribution are more abundant than in a normal distribution. For the PCB congeners discussed here, the kurtosises of congeners 19, 100, and 201 are relatively high, indicating that there is some censoring for these distributions. The question now becomes whether these distributions can be corrected for this censoring.

Let us assume that the PCB concentrations are detected with a decreasing probability below their limit of quantitation (LOQ) and that this probability is not zero. In other words, the LOQ does not represent a step function below which the compounds are not detected. I parametrized this idea with the following probability function:

$$P(x_i) = 1 \text{ for } x_i > c_3$$
$$P(x_i) = (c_3 - x_i + 1)^{-1} \text{ for } x_i \le c_3 \quad (4)$$

where x and c_3 are logarithmically transformed concentrations. In this case, c_3 is the logarithm of the LOQ. This function says that the probability of measuring a given concentration is unity for concentrations above the LOQ, but this probability decreases as the concentration gets smaller relative to the LOQ, in this case given by $c_3 - x_i$. The problem is that I do not know the true value of c_3, which will be different for each PCB congener. I can, however, find this value by least-squares curve fitting the function

$$f(x_i) = P(x_i)u(x_i) \quad (5)$$

where $P(x_i)$ is given by eq 4 and $u(x_i)$ is the uncensored distribution function given by

$$u(x_i) = c_0 \exp[-(x_i - c_1)^2/2c_2^2] \quad (6)$$

where c_0 is a scale factor, c_1 is the mean of the uncensored distribution, and c_2 is its standard deviation. This equation represents an assumption. It assumes that the data that have been censored and that we cannot see are, in fact, log-normally distributed. On the basis of Figure 1, this seems like an acceptable assumption. There is an additional constraint: The integral of eq 6 must equal the total number of possible measurements (N), or because these are discrete numbers

DOI: 10.1021/acs.estlett.5b00223
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

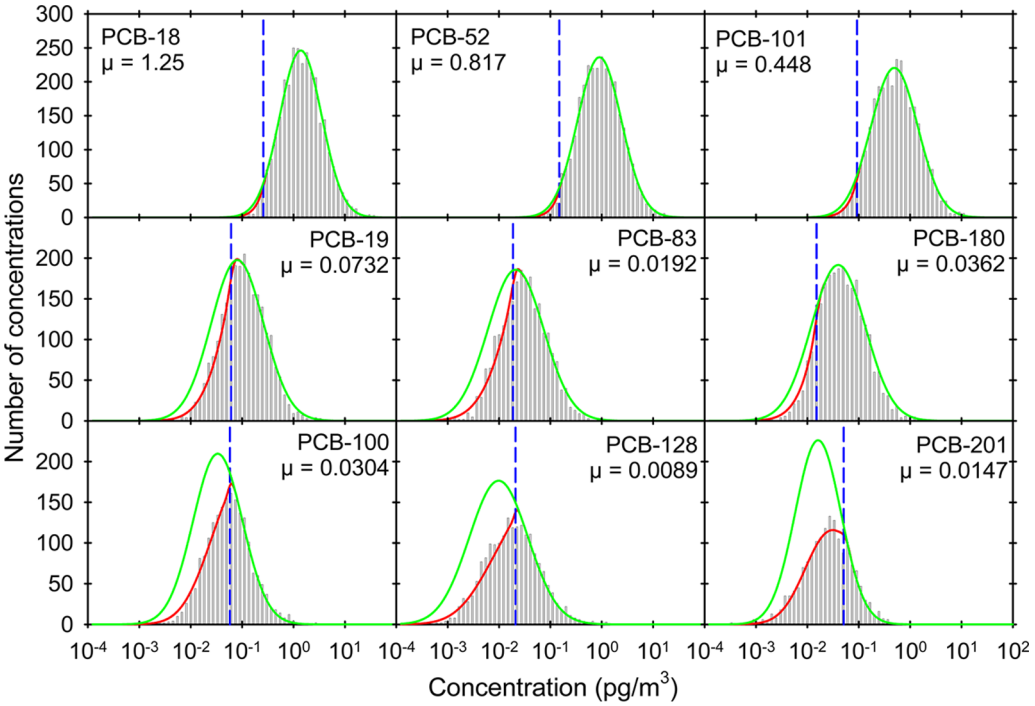


Figure 2. Histograms as shown in Figure 1. The red line was fitted using eq 5; the green line is the uncensored distribution function fitted with eq 6, and the blue dashed line is the limit of quantitation determined using eq 4. The geometric means (μ), after correction for censoring, are given for each congener. The parameters of these fits are listed in Table S3.

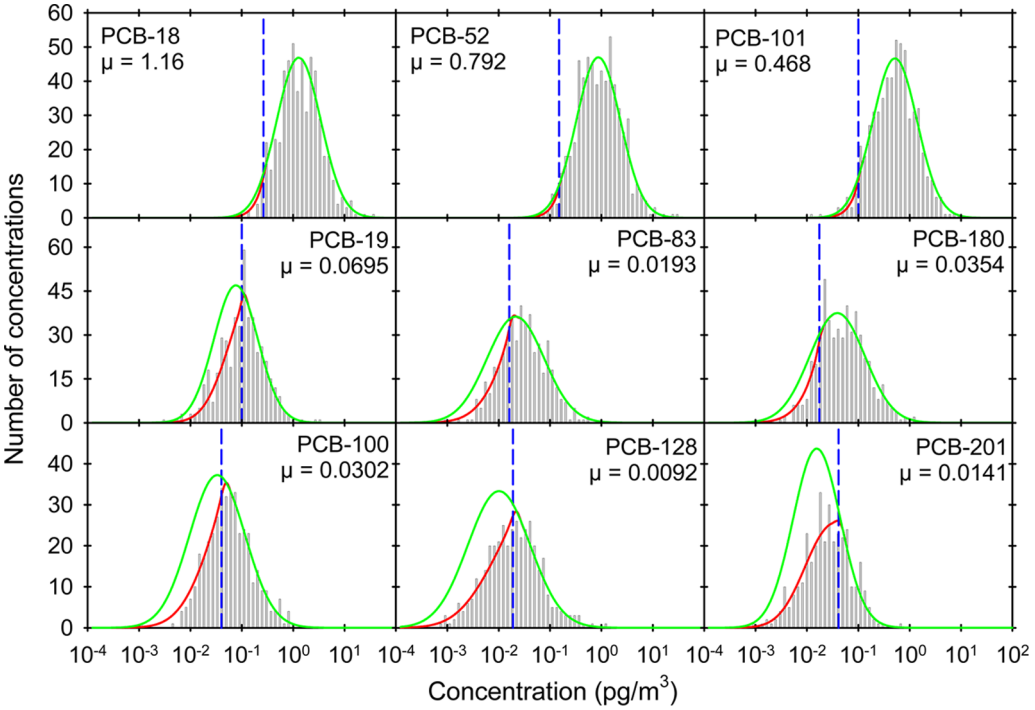


Figure 3. Histograms of 20% randomly selected PCB concentrations for each congener. The axes are the same as Figure 1. The meanings of the red, green, and blue lines are the same as in Figure 2. The geometric means (μ), after correction for censoring, are given for each congener. The parameters of these fits are listed in Table S4.

$$N = \sum_{i=1}^{90} u(x_i) \tag{7}$$

The actual curve fitting was done in an Excel spreadsheet using the Solver feature. This is a nonlinear curve fit, so it is beneficial to have guesses of the starting values of the four constants in eqs 4 and 6. The initial estimates of c_1 and c_2 were from the descriptive statistics listed in Table S1; the initial estimate of c_0 was the maximum of the counts in the histogram, and the initial estimate of c_3 was simply c_1. As an example of the actual curve fitting process, an operating spreadsheet for the PCB-128 histogram data is given in the Supporting

DOI: 10.1021/acs.estlett.5b00223
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

Information, where *prob* is the probability from eq 4, *base* is the normal distribution from eq 6, *expect* is the expected value from eq 5, and *residsq* is the squared difference betweeen the measured count and the expected value. The sum of these *residsq* values [the sum of squares (SOS)] is the cell to be minimized using Solver, which systematically varies the values of c_0 to c_3 to minimize the SOS, while forcing the total of the *base* column (eq 7) to be the total number of possible measurements. In Solver, this is accomplished by setting a constraint such that the cell labeled *count* must be 2909, in this case. The result of this calculation is a least-squares fit of the data, and this fit is shown in the Supporting Information for PCB-128. This spreadsheet also shows the resulting geometric mean [after correction (see eq 3)], the LOQ [after correction (see eq 3)], the mean sum of squares, and the calculated percent detected.

Using this approach, all of the histograms for the nine congeners were fitted using eq 5 with the constraint that N = 2909. These results are plotted in Figure 2, and the resulting parameters are listed in Table S3. The LOQs determined with this approach are also shown in Figure 2. The geometric means and calculated LOQ values are summarized in columns 6 and 8 of Table 1. Interestingly, even for congeners 18, 52, and 101, which were measured in almost every sample, the LOQs are in the range of 0.1−0.3 pg/m^3, indicating that even for these PCB congeners there is some underestimation of their lowest concentrations. For congeners 100, 128, and 201, which were measured in only approximately two-thirds of the samples, the LOQs are higher than the geometric means and the medians of the measured concentrations (compare columns 3 or 4 and 8 of Table 1). As expected, the means are about the same as previously calculated from eq 2 for PCB congeners 18, 52, and 101, but they are much lower for the congeners with left-censored data (compare columns 5 and 6 of Table 1). On average, the geometric means are ∼20% lower for congeners 19, 83, and 180 and ∼40% lower for congeners 100, 128, and 201. These are large differences and indicate the importance of correcting left-censored data. It is interesting that these percent differences are about the same as the percent of missing values. The calculated LOQ values for the censored distributions are on the order of 0.050 pg/m^3, which are much higher than the analytical detection limit of 0.001 pg/m^3, indicating that the analytical detection limit is not always a good measurement of the overall method's sensitivity.

The data set I have investigated here is large, consisting of up to 2909 measurements. One wonders if the methods used here would be as effective if there were fewer data. To investigate the sensitivity of the method to the number of data, I randomly selected 20% of the data from the full data set and repeated the calculations described above. Of course in this case, N = 582. The results are shown in Figure 3, and all of the fitted parameters are listed in Table S4. As expected, the histograms are noisier, but the resulting geometric means and LOQ values are similar to those found using the full data set (compare columns 6 and 7 and columns 8 and 9 of Table 1). This indicates that the approach described here is statistically robust and could be applied to smaller sets of data.

This work suggests several conclusions. First, these PCB concentrations are log-normally distributed. This conclusion is based on several observations. (a) The fitted curves shown in Figure 1 give means that are virtually the same as the geometric means and the medians of the data itself. (b) The skewness and kurtosis values are, for the most part, near zero, as expected for a normal distribution. (c) The mean sums of squares (SOS) are small. (d) Visually, the red fitted curves in Figure 1 match the data. This conclusion indicates that the geometric mean or median concentration is the best measure of the central tendency for those PCB congeners detected in almost all of the samples. Second, for those PCB congeners detected in some of the samples, the approach described here provides a good and statistically robust estimate of the true geometric mean of the distribution, which is lower than the descriptive statistics or a simple fit with a normal distribution would indicate. This approach also gives an estimate of the limit of quantitation (LOQ) of each PCB congener, values of which are different for each congener. Third, it remains to be seen if this approach can be universally applied, particularly to data that are more severely censored.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.estlett.5b00223.

Tables S1−S4 giving the descriptive statistics and the regression results related to Figures 1−3 (PDF)

Working Excel spreadsheet demonstrating the calculations for PCB-128 (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*Telephone: 812-855-0193. E-mail: hitesr@indiana.edu.

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

I thank Amina Salamova for helpful comments. This work was supported by the Great Lakes National Protection Office of the U.S. Environmental Protection Agency through cooperative agreement GL00E01422 (Todd Nettesheim, project officer), for which I am grateful.

■ REFERENCES

(1) Limpert, E.; Stahel, W. A.; Abbt, M. Log-normal distributions across the sciences: Keys and clues. *BioScience* **2001**, *51*, 341−352.

(2) Succop, P. A.; Clark, S.; Chen, M.; Galke, W. Imputation of data values that are less than a detection limit. *J. Occup. Environ. Hyg.* **2004**, *1*, 436−441.

(3) Helsel, D. R. *Statistics for Censored Environmental Data using Minitab® and R*; Wiley: Hoboken, NJ, 2012.

(4) Carlson, D. L.; Hites, R. A. Temperature dependence of atmospheric PCB concentrations. *Environ. Sci. Technol.* **2005**, *39*, 740−747.

(5) Salamova, A.; Venier, M.; Hites, R. A. Revised temporal trends of persistent organic pollutant concentrations in air around the Great Lakes. *Environ. Sci. Technol. Lett.* **2015**, *2*, 20−25.

(6) Wu, R.; Backus, S.; Basu, I.; Blanchard, P.; Brice, K. A.; Dryfhout-Clark, H.; Fowlie, P.; Hulting, M. L.; Hites, R. A. Findings from quality assurance activities of the Integrated Atmospheric Deposition Network. *J. Environ. Monit.* **2009**, *11*, 277−296.

(7) Venier, M.; Hites, R. A. Time trend analysis of atmospheric POPs concentrations near the Great Lakes since 1990. *Environ. Sci. Technol.* **2010**, *44*, 8050−8055.

DOI: 10.1021/acs.estlett.5b00223
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX