Environmental Science: Atmospheres

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5ea00005j

A novel statistical approach for analyzing environmental pollutant data with detection limits: atmospheric organochloride pesticide concentrations near Tibet's Namco Lake as a case study†‡

Lidong Huang, *[a] Qian Zheng,[a] Hongyang Wang[b] and Daquan Sun[c]

In the analysis of pollutant data, concentrations below the analytical detection limit are commonly handled by substituting a constant value between zero and the limit of detection (LOD). However, this substitution can introduce significant bias under certain conditions. To address this issue, we have derived weight expressions that eliminate bias for lognormal and gamma data. These weights, applied to LOD/2 substitutions, can be calculated using available ranges of means, standard deviations and censoring proportions. We evaluated the performance of our weighted substitution (ωLOD/2) method using both simulated datasets with censoring proportions ranging from 5% to 50% and actual atmospheric α-HCH and HCB data from Tibet's Namco Lake. The ωLOD/2 method was compared against LOD/2 substitution, maximum likelihood estimation (MLE), and regression on order statistics (ROS). The results demonstrate that with small sample sizes (<160), although MLE and ROS did not show larger bias, ωLOD/2 outperforms both methods in estimating arithmetic and geometric means in most scenarios. It is also worth noting that ROS is currently limited to estimating summary statistics under the assumption of a lognormal distribution and cannot be applied to gamma-distributed data. In addition, ωLOD/2 provides standard deviation estimates comparable to those from MLE, with biases remaining within 5% in the majority of cases. Therefore, the proposed method is particularly suitable for situations involving small sample sizes. The application of our method to six censored atmospheric organochloride pesticide concentrations from Namco Lake further highlights its advantages in practical settings. To facilitate easy adoption by researchers, a free web app was developed that integrates our proposed weighting method with censored data distribution fitting.

Received 9th January 2025
Accepted 17th June 2025

DOI: 10.1039/d5ea00005j

rsc.li/esatmospheres

Environmental significance

Accurate estimation of atmospheric pollutants is vital for assessing environmental risks and implementing effective protection measures. Traditional methods for handling data below detection limits often introduce bias, leading to potentially misleading conclusions about pollutant levels. Our study presents a weighted substitution method (ωLOD/2) that significantly improves the accuracy of pollutant estimates, particularly for lognormal and gamma distributions. Applied to data from Tibet's Namco Lake, this method ensures more reliable environmental assessments. By enhancing the precision of pollutant monitoring, our approach supports better-informed environmental policies and strategies, ultimately contributing to more effective environmental protection efforts.

[a]College of Resources and Environmental Sciences, Inner Mongolia Agricultural University, Key Laboratory of Agricultural Ecological Security and Green Development at Universities of Inner Mongolia Autonomous, Inner Mongolia Key Laboratory of Soil Quality and Nutrient Resource, Hohhot 010018, China. E-mail: ldhuangnz@163.com

[b]Litian Statistics Studio, Nanjing 210044, China

[c]Institute of Microbiology, Czech Academy of Sciences, Vídeňská, 1083, Praha 4, 14220, Czech Republic

† The original R code for the web app is archived on GitHub (https://github.com/Lidong-Huang/Censored.DF/tree/main).

‡ Electronic supplementary information (ESI) available: The mathematical processes for computing conditional population statistics. Fig. S1 and S8 illustrate the geometric principle of the LOD/2 substitution method. Fig. S2 and Table S2 present the parameter settings for data simulations. Fig. S3 shows the distribution fitting results for censored OCPs. Fig. S4, S5 and Tables S3–S6 demonstrate the influence of parameters on the accuracy of the LOD/2 substitution. Fig. S6 and S7 depict the nonlinear relationship between the known information of censored data and weights. Fig. S9 highlights the difference between the geometric mean and median for gamma data. Fig. S10–S12 compare the accuracy of different methods. Fig. S13 provides a screenshot of the web app. Table S1 lists the formulae used in this study, and Table S7 summarizes the weight (ω). See DOI: https://doi.org/10.1039/d5ea00005j

1. Introduction

Organochlorine pesticides (OCPs) pose significant risks to both human health and ecological systems. These compounds are highly persistent in the environment, bioaccumulate in the food chain, and exhibit long-term toxicity. For humans, exposure to OCPs can lead to serious health issues, including cancer, endocrine disruption, reproductive disorders, and neurological damage.[1,2] They can enter the body through contaminated food, water, and air and are stored in fatty tissues, leading to prolonged exposure. OCPs are transported by air through volatilization from contaminated surfaces and soils.[3] Once airborne, they can travel long distances *via* atmospheric currents, eventually depositing in remote areas through dry deposition and precipitation, leading to widespread environmental contamination.

Tibet, known as the "Third Pole" and the "Water Tower of Asia," plays a crucial role in the region's environmental health. The surrounding countries' industrial activities have led to significant organic pollution, including OCPs. These pollutants are transported through atmospheric currents and eventually settle in Tibet's pristine environment. Monitoring OCP levels in Tibet is essential not only for protecting its unique ecosystem and public health but also for understanding the broader impacts of transboundary pollution. Previous studies showed seasonal variations in OCPs in the Tibet Namco Lake atmosphere, with some compounds falling below limits of detection (LOD).[4,5] This complicates accurate monitoring and risk assessment, as low detection limits hinder the ability to quantify exposure levels, evaluate long-term environmental impacts, and develop effective regulatory policies for public health protection. In the same region, while most OCPs are below detection limits in lake water, they are fully detected in fish.[5] This suggests that low environmental levels of OCPs do not necessarily equate to low risk. Even trace pollutants can bioaccumulate in organisms through biomagnification, posing significant ecological and health risks. Therefore, accurately estimating the statistics of datasets with non-detectable concentrations is crucial for understanding pollution characteristics and assessing risks. Such data, where some measurements fall below the detection limit, are commonly referred to as left-censored datasets. A diverse array of methods have been developed to cope with this problem caused by censored data, such as substitution $(\mathrm{LOD}/2, \mathrm{LOD}/\sqrt{2})$,[6,7] the maximum likelihood method (MLE),[6,8,9] regression on order statistics (ROS),[6,10] and the fitting distribution curve method.[11] Besides, Tobit models, originally developed in econometrics, are designed to address left-censored dependent variables by modeling an underlying latent variable through MLE.[12,13] The Tobit model is particularly valuable when the goal is regression-based inference.[14] However, the Tobit model requires strong assumptions—most notably, that the latent variable follows a normal distribution, which may not hold in environmental datasets where variables often follow gamma distributions.[15] Additionally, Tobit models are primarily designed for estimating regression coefficients, rather than directly estimating summary statistics such as means and standard deviations, which are the focus of our study. Survival analysis, particularly the Cox proportional hazards model and Kaplan–Meier estimator, has long been applied to right-censoring and more complex censoring structures.[6,16,17] The Cox proportional hazards model is a widely used method in survival analysis for modeling the relationship between the time until an event occurs and one or more predictor variables.[18] Similar to Tobit models, the Cox proportional hazards model is primarily designed for regression analysis involving covariates, rather than for estimating summary statistics such as the mean or standard deviation. In addition, existing literature suggests that the Kaplan–Meier estimator may be less accurate than MLE or ROS when estimating distributional summary statistics in environmental datasets.[6,19]

Among these methods, substitution is widely used, but it remains a subject of considerable debate.[20] In some scenarios, the substitution method performs exceptionally well. According to George *et al.*,[21] "Threshold/2 substitution was the least biased" method for determining the concentration means and standard deviations (sd) of dibenzo[*a*,*h*]anthracene in stove chimney emissions. Furthermore, many studies have concluded that substitution is sometimes comparable to other methods in estimating means.[7,16] When discussing which method is recommended for handling censored data, Hites commented "Clearly, as long as more than half of the data are above the LOD, using the median or geometric mean with the <LOD or missing values replaced by LOD/2 causes little bias". Besides, to our knowledge, the U.S. CDC uses a method where concentrations below the LOD are assigned a value equal to the $\mathrm{LOD}/\sqrt{2}$ for calculating geometric means in studies on human exposure to environmental chemicals.[22]

Simultaneously, we must also acknowledge that the substitution method is not indefensible. For example, it tends to be less accurate than MLE and ROS in estimating the population mean.[6] Moreover, the common choices for substitution are LOD/2 or $\mathrm{LOD}/\sqrt{2}$, while the results strongly depend on the substituted values. The geometric diagram of the principle of substitution (Fig. S1) indicates that several factors might influence the accuracy of substitution: (1) sample size, which determines the smoothness and step length of the Empirical Cumulative Distribution Function (ECDF) curve; (2) the percent of observations below the LOD (<LOD%) and (3) distribution parameters, which affect the steepness and shape of the ECDF curve. LOD/2 substitution assumes that all censored values are equal to the midpoint, which ignores the true variability of values below the detection limit. Inspired by this question, we propose using a weighted substitution value that ensures consistent accuracy. Our proposed method is based on the idea that the estimation of summary statistics for left-censored values can be informed by the observed portion of the data, due to the inherent structure of the dataset. —much like reconstructing missing sections of a broken picture based on the remaining visible parts.

Generally, it is known that concentrations of pollutants are skewedly distributed.[6,23,24] The vast majority of environmental

scientists assume that pollutant distributions follow a lognormal pattern.[6,7,24–26] Under such an assumption, it is suggested that the geometric mean (gm) should represent central tendency, as the median equals gm for lognormal data (Table S1‡).[7,16,24,27] However, we found that more than half of OCPs in the atmosphere of Namco Lake followed a gamma distribution (Fig. S2‡). Notably, the median of gamma data does not align with the geometric mean (Table S1‡). This introduces a pertinent question: in cases where the gamma distribution better fits the data, should one use the arithmetic mean (am) or gm to represent data? Additionally, many studies overlook the crucial aspect of justifying whether censored data actually follow a lognormal or gamma distribution. The lack of this justified process can potentially lead to misleading conclusions. In this context, EnvStats is an *R* package that provides a wide range of tools for analyzing censored data, estimating distribution parameters, conducting goodness-of-fit tests, and generating visual diagnostics.[23] It includes functions such as elnormCensored() and egammaCensored(), which implement MLE for left-censored data under lognormal and gamma distributions. However, MLE can exhibit greater bias than substitution methods when applied to small sample sizes, as it relies on asymptotic properties that may not hold in such cases.[19,28] To facilitate the process and simplify the coding work, A user-friendly fitting tool would enable more informed decision-making and enhance the reliability of environmental studies.

In summary, the objectives of this study are to (i) explore the factors impacting the accuracy of LOD/2 substitution; (ii) develop a more accurate substitution method by weighting LOD/2. We approximate the weight through a function of the following form:

Weight $\sim f(<\text{LOD}\%, E(X|X > \text{LOD}), \text{SD}(X|X > \text{LOD}))$; (iii) apply this improved method to deal with censored OCPs in the Tibet Namco Lake atmosphere.

2. Materials and methods

We evaluated the bias associated with left-censoring and the subsequent statistical analysis of concentration measurements through both case and simulation studies. The case study utilized OCP concentration measurements, while the simulation study employed both lognormal and gamma distributions similar to the observed OCP distribution.

2.1 Case study

Namco Lake ((30°30′–30°56′ N, 90°16′–91°01′ E), situated in the central Tibet Autonomous Region, is the second-largest lake in Tibet and the third-largest saltwater lake in China. Namco is considered the highest large lake in the world and is considered one of the regions least impacted by OCPs. The data in our case study were from an air pollutant monitoring experiment at the Namco Monitoring and Research Station for Multisphere Interactions.[5] A total of 47 air samples were collected from September 2012 to September 2014. Eight OCPs (α-HCH, β-HCH, γ-HCH, HCB, *o*,*p*′-DDE, *p*,*p*′-DDE, *o*,*p*′-DDT, and *p*,*p*′-DDT) in the air were determined using gas chromatography–mass spectrometry (GC-MS). The LOD were derived as the mean blank concentration plus three times its standard deviation, based on 600 m^3 air samples.[5] We found that 6 out of the 8 OCPs had concentrations below the LOD. Therefore, OCPs are the analytes needed for our assessment of bias when calculating the sample statistics.

2.2 Simulation study

Lognormal data with parameters μ log and σ log and gamma with shape (α) and rate (β) are simulated. Factors influencing the data structure, including sample size (n), μ log, σ log, α, β and <LOD% are listed in Fig. S2 and Table S2.‡ The ranges of the factors are referred to in publications.[6,8,27] μ log, σ log, α and β are linked by equal expectation and variance. Specifically, the expectation (μ_L) and variance (σ_L^2) for the lognormal distribution are

$$\mu_L = e^{\mu \log + \frac{\sigma \log^2}{2}} \cdot {\sigma_L}^2 = e^{2\mu \log + \sigma \log^2}\left[e^{\sigma \log^2} - 1\right] \quad (1)$$

And expectation (μ_G) and variance (σ_G^2) for the gamma distribution are

$$\mu_G = \frac{\alpha}{\beta} \cdot {\sigma_G}^2 = \frac{\alpha}{\beta^2} \quad (2)$$

Based on $\mu_L = \mu_G$ and ${\sigma_L}^2 = {\sigma_G}^2$, once we set the parameters for the lognormal distribution, the parameters for gamma are also fixed. The true sample mean and sd for the complete data are known. After artificially censoring, we assessed the bias in the mean and sd estimation at each censoring level. We used the relative bias (RB) to compare the accuracy of different statistical methods:

$$\text{RB}(\%) = (\text{estimated} - \text{true})/\text{true} \times 100 \quad (3)$$

2.3 Procedures to get the substitution weight

For simulated data, LOD is virtually set. In this case, the data below the LOD is also known. Assuming that the simulated pollutant X has n observations, $X = \{x_1, x_2, \ldots, x_n\}$, the true am of X could be expressed as follows:

$$am = \frac{n_< \times \overline{x_<} + n_> \times \overline{x_>}}{n} \quad (4)$$

where $\overline{x_<}$ is the conditional mean of <LOD X. $n_>$ is the sample size of >LOD X. $\overline{x_>}$ is the conditional mean of >LOD X. Now, the <LOD X array was substituted by '$\omega \times \text{LOD}/2$' and ω is weight. The am_ω can be as follows:

$$am_\omega = \frac{n_< \times \omega\frac{LOD}{2} + n_> \times \overline{x_>}}{n} \quad (5)$$

Let the estimated am_ω equal the true am (assuming 0% bias), formula (4) = formula (5), then we can easily derive the ω:

$$\omega = 2\frac{\overline{x_<}}{\text{LOD}} \quad (6)$$

Likewise, we could also get the weight for calculating gm, ω':

$$\omega' = 2\frac{\overline{x_{g<}}}{\mathrm{LOD}} \text{where is conditional gm of } \overline{x_{g<}}\ X < \mathrm{LOD} \quad (7)$$

Obviously, weight computation needs the conditional mean of <LOD X, which is unrealistic for actual measured censored data, because no one knows the value. To address this issue, our approach involves constructing a range of weight variations using simulated data and leveraging the available information from censored data to predict the weights.

2.4 Model weight with available information of the censored data

As we know, sample $\overline{x_<}$ (or $\overline{x_{g<}}$) will asymptotically approach conditional population expectation $E(X|X < c)$ or $E[\mathrm{gm}|X < c]$ (c is LOD) and we can get

$$\omega \approx 2\frac{E(X|X<c)}{c} \text{or } \omega' \approx 2\frac{E[\mathrm{gm}|X<c]}{c} \quad (8)$$

The detailed mathematical derivation process of the conditional expectation for the lognormal and gamma distribution can be found in the ESI,‡ and we also summarized the related formulae in Table S1.‡

Utilizing the mathematical expressions for $E(X|X < c)$ and $E[\mathrm{gm}|X < c]$ (Table S1‡), we can easily compute the ω (ω') based on formula (8). For censored data, the known information includes the proportion below the detection limit (<LOD%) and the mean and standard deviation of the data above the LOD. So, we tried to explore the relationships between ω (ω') and $E(X|X \geq c)$ and $\mathrm{sd}(X|X \geq c)$ at different censored levels. And the expressions of $E(X|X \geq c)$ and $\mathrm{sd}(X|X \geq c)$ for lognormal and gamma are derived in the ESI,‡ and in our previous paper.[8] After extensive trials, the following nested models were found to be the best to fit $\hat{\omega}$

$$\widehat{\omega} = \begin{cases} a \cdot t^b & \text{Lognormal data, } a, b \text{ are constants, } t = \dfrac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)} \\ & \text{predict } a \text{ or } b \text{ with linear equation by } < \mathrm{LOD}\% \\ a \cdot e^{b \cdot t} & \text{Gamma data, } a, b \text{ are constants, } t = \dfrac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)} \\ & \text{predict } a \text{ or } b \text{ with quadratic equation by } < \mathrm{LOD}\% \end{cases} \quad (9)$$

2.5 Comparison of the performance of the weighted substitution with that of other methods

α-HCH and HCB and simulated data were used to compare the performance of the weighted method with that of the most sophisticated methods. Means and sd were calculated after censoring by using the following approaches: LOD/2 substitution, the weighted method (ωLOD/2), MLE and ROS. The likelihood function of MLE for lognormal or gamma-censored data is expressed as follows:

$$L(\theta) = \prod_{i \in D} F_X(\theta|L_x) \prod_{i \in M} f_X(\theta|x_i)$$

where θ is a parameter. F_X is the cumulative distribution function (CDF) and f_X is the probability density function (PDF). L_x is the LOD of x. D is the vector of below LOD. M is the vector of observed x.

And $\widehat{\theta} = \mathrm{argmax} - \ln L(\theta)$. The corresponding R code is listed in the ESI.‡[θ]

Kaplan–Meier and non-parametric quantile methods were excluded because of poor performance.[6,19] Referring to George *et al.*,[6] simulated data values were generated for 1000 data sets with $n = 50$ for each censoring level from two lognormal and two gamma distributions, representing two skewed levels and RBs are averaged.

Since both α-HCH and HCB have been completely detected, their statistics, such as the am, gm and sd, are known. Biases of different estimation methods were compared by artificially censoring the levels at 10%, 30%, and 50%. MLE and ROS estimates are computed based on the distribution test of α-HCH and HCB. ROS is only for censoring data where lognormal distribution is assumed.

2.6 Statistical analysis

Data simulations and analysis are completed through *R* language (R 4.3.2).[29] Lognormal data simulation was performed using the *R* function *rlnorm*(n, μ log, σ log), and gamma data was simulated using the function *r*gamma(n, α, β). When using LOD/2 substitution, we employed a linear model lm (…) function to test whether the factors had a significant ($P < 0.05$) effect on the RB of the am or gm estimation. In the model, the log-transformed RB served as the dependent variable, while μ log, σ log (or α, β), n and <LOD% were treated as independent variables. Standardized coefficients were extracted using the lm.beta (…) function in *R*. Weights ω were computed based on lognormal and gamma data simulated with μ log, σ log, α and β listed in Table S2.‡ In the simulated data, LOD was artificially set and <LOD% was obtained. $\mathrm{sd}(X|X \geq c)$ and $E(X|X \geq c)$ could be computed accordingly (Table S1‡).

Non-linear relationship between ω and $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$, as shown in formula (9), was fitted using the *R* nls (…) function. Consequently, the parameters a and b were obtained at different <LOD% levels. Linear and quadratic functions were applied to fit the relationships between a (b) and <LOD% for lognormal and gamma distribution separately. To test the goodness of fit of the nested models $R^2 = 1 - \frac{\sum(\omega_i - \hat{\omega}_i)^2}{\sum(\omega_i - \overline{\omega})^2}$ was calculated accordingly. Given the satisfaction of formula (9), $\omega\hat{\omega}$ could be computed using the available information of the censored data, specifically the <LOD%, sd, and mean of the detected observations. The sample statistics of weighted substitution were then computed as follows:

$$\mathrm{am}_{\omega\mathrm{LOD}/2} = \frac{n_< \times \hat{\omega}\frac{\mathrm{LOD}}{2} + n_> \times \overline{x_>}}{n};$$

$$\mathrm{gm}_{\omega\mathrm{LOD}/2} = \exp\left(\frac{n_< \times \log\left(\hat{\omega}\frac{LOD}{2}\right) + n_> \times \overline{\log(x_>)}}{n}\right); \text{ where}$$

$\overline{\log(x_>)}$ is the mean of logarithm x, which is detected.

$$\mathrm{sd}_{\omega\mathrm{LOD}/2} = \left(\frac{n_< \times \left(\hat{\omega}\frac{LOD}{2} - \mathrm{am}_{\omega LOD/2}\right)^2 + \sum_{(n-n_<+1)}^{n}\left(x - \mathrm{am}_{\omega LOD/2}\right)^2}{n-1}\right)^{0.5}$$

R mle (…) and *r*os (…) functions were applied to implement the MLE and ROS estimation. lognormal and gamma distributions were considered candidates for best-fitting parametric distributions for GC-MS-determined OCP measurements. However, it is well known that testing the distributions of censored data, compared to completely observed data, is challenging.[6,30] To simplify the process of fitting distributions of OCPs, we have developed a web app using the *R* Shiny package to censor environmental data. This app is available at **https://lidonghuang.shinyapps.io/Censored_fitting_weighting_substitution/**. In the app, the function *gofTestCensored* (…) from package 'EnvStats' is used to perform goodness-of-fit tests. These tools help visually and statistically assess how well the chosen distribution fits the observed data. The Shapiro–Wilk test, ECDF plot and QQ plot were used to evaluate and indicate the better distribution between the candidates, lognormal and gamma. To enable other scientists to conveniently utilize the weighted substitution method proposed in this study, we have integrated it into the web app as well. Instructions on how to use the app are provided on the home page of the app.

3. Results

3.1 Distributions of air OCP measurements

Fig. 1 illustrates the concentration ranges of the eight atmospheric OCPs. Notably, two were detected at full prevalence (α-HCH and HCB), while the remaining six displayed varying

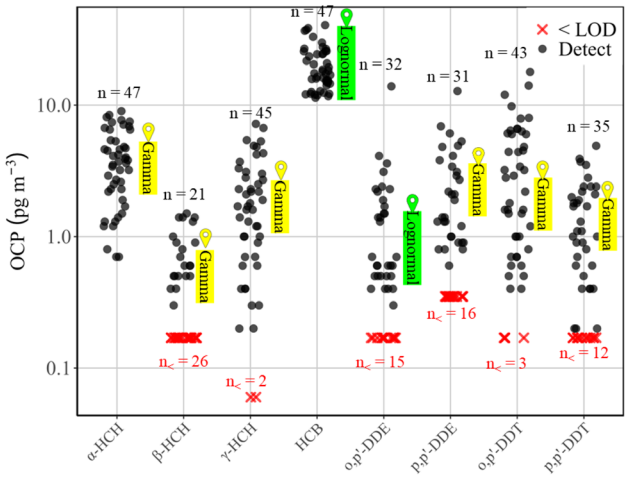

Fig. 1 Concentrations of OCPs in the Tibet Namco lake atmosphere. The distribution fittings were performed using a web app developed in this paper (Fig. S3‡).

degrees of concentrations <LOD. For these six OCPs, the <LOD% ranged from 9% (*o*,*p*′-DDT) to 60% (β-HCH). The Shapiro–Wilk test and QQ plot with fitted lognormal and gamma distributions indicate gamma to be the best fitting for these six. This conclusion is supported by the higher *p*-values and the points lying closer to the straight line in the QQ plot, which both suggest a better fit (Fig. S3‡). Therefore, it may be prudent to question the assumption that pollutant concentrations merely follow a lognormal distribution. If pollutants follow a gamma distribution, it necessitates a re-evaluation of several theoretical frameworks. For example, the selection of the likelihood function for the MLE, the suitability of the ROS method, and the representativeness of the gm will all require reconsideration. At present, estimates of the am, gm and sd of these six censored pollutants were required on the basis of robust methods.

3.2 The RB of mean estimation by LOD/2 substitution for simulated data

As illustrated in Fig. 2, for the lognormal type, the RB of am and gm estimated by LOD/2 substitution ranged from −14% to 2% and from −21% to 19% respectively. For the gamma type, the RB of am and gm estimated by LOD/2 substitution ranged from −13% to 6% and from −21% to 842%. The accuracy of both am and gm is significantly influenced by skewness of data, as skewness of lognormal data is given by $[e^{\sigma\log^2} + 2]\sqrt{[e^{\sigma\log^2} - 1]}$ and skewness of gamma data is $2/\sqrt{\alpha}$ (Tables S3–S6‡). Higher skewness and <LOD% lead to greater uncertainty in mean estimation (Fig. 1). Other factors, such as μ log, β and n, had little effect on the RB of both am and gm (Fig S4–S5‡). This result can explain why some studies find the LOD/2 substitution method to have smaller bias, while others find it to have larger bias,[6,7,19] This discrepancy is evidently related to the distribution structure of the data (Fig. S1‡).

3.3 Weighted LOD/2 substitution

When adjusting the LOD/2 to correct bias in mean estimation using simulated data, the averaged weights used to compute am

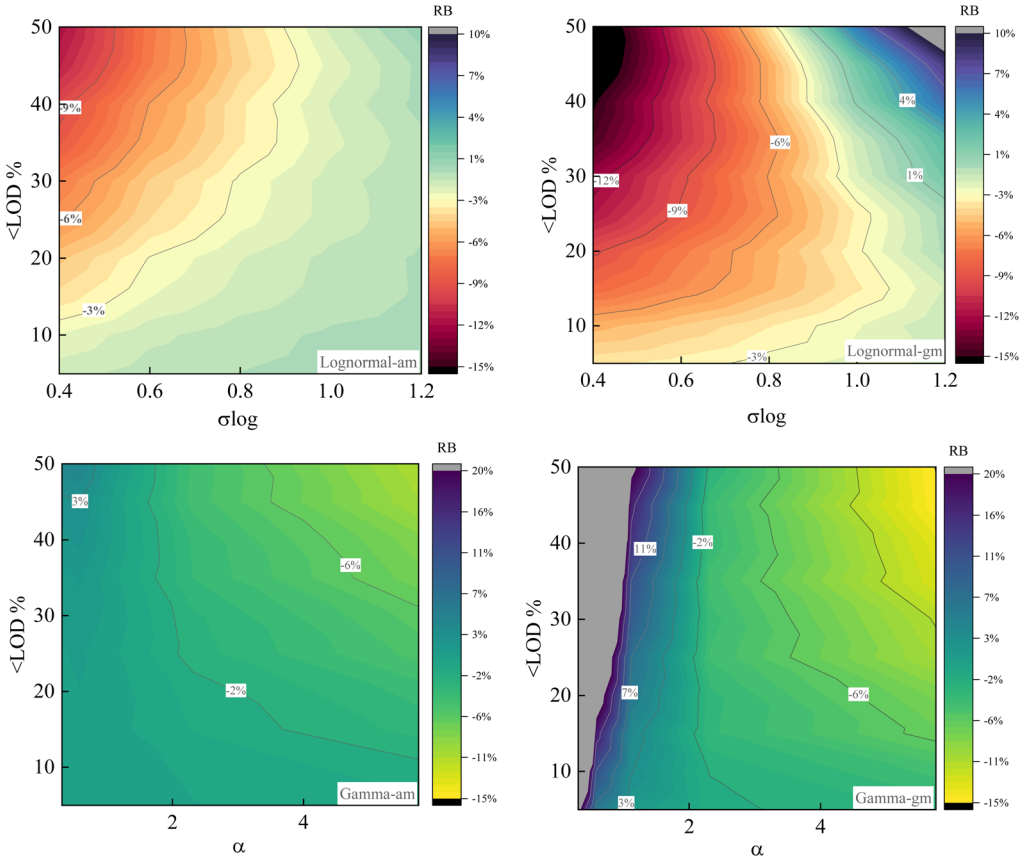

Fig. 2 Response of Relative Bias (RB) for arithmetic mean (am, left) and geometric mean (gm, right) in simulated censored data: Effects of lognormal σlog and <LOD% (top) and gamma shape parameter α and <LOD% (bottom), with different scales applied.

and gm for lognormal censored data were 1.33 and 1.24, and 0.98 and 0.76 for gamma censored data respectively (Table S7‡). The relationship between the weights and known characteristics of the censored data was illustrated in scatter plots (Fig. 3). A general trend observed is that the weight decreased with an increase in the ratio of sd to the expectation of observations > LOD, $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$, and decreased with <LOD%. Notably, the weights for gamma data were nearly constant at the highest $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$ stage, suggesting that they are more affected by <LOD%. Previous study pointed out that the reasonable limit for $\frac{\mathrm{sd}(X)}{E(X)}$ is 0.1–2,[31] which is consistent with the range in this study, indicating that our weights can be applied to the majority of cases.

Aiming to predict weights using available information $\left(\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}, <\mathrm{LOD}\%\right)$, various conventional nonlinear models are used to fit the relationship between ω and $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$ at fixed < LOD%. The power function and exponential functions as shown in formula (9) are found best to predict ω for lognormal and gamma distribution (Fig. S6 and S7‡). To embrace the <LOD% in the prediction, constants of formula (9) are linked with <LOD% using linear and quadratic equations for lognormal and gamma types respectively (Fig. S6 and S7‡). The full models (incorporating both $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$ and <LOD%) are summarized in Table 1. The selected model shows a better goodness of fit to the weights with an $R^2 > 0.98$.

4. Discussion and applications

The accuracy of LOD/2 substitution for estimating am or gm was markedly affected by skewness and <LOD% (Fig. 1), which can be explained by the geometric area demonstration (Fig. S8‡). Essentially, high skewness results in a steep initial stage of the ECDF curve, increasing the bias of LOD/2 substitution. The bias of LOD/2 substitution is magnified for gm due to the transformation through an exponential algorithm (Fig. S2‡). Apparently, large <LOD% can lead to unacceptable accuracy levels, a concern also highlighted by others.[6,7,32]

4.1 Differences in representativeness of gm between lognormal and gamma data

Some researchers suggest using gm instead of am to represent the central tendency of lognormal censored data,[7,16,32] as supported by the fact that the median of lognormal data equals the gm (Table S1‡). Although the gm coincides with the median for

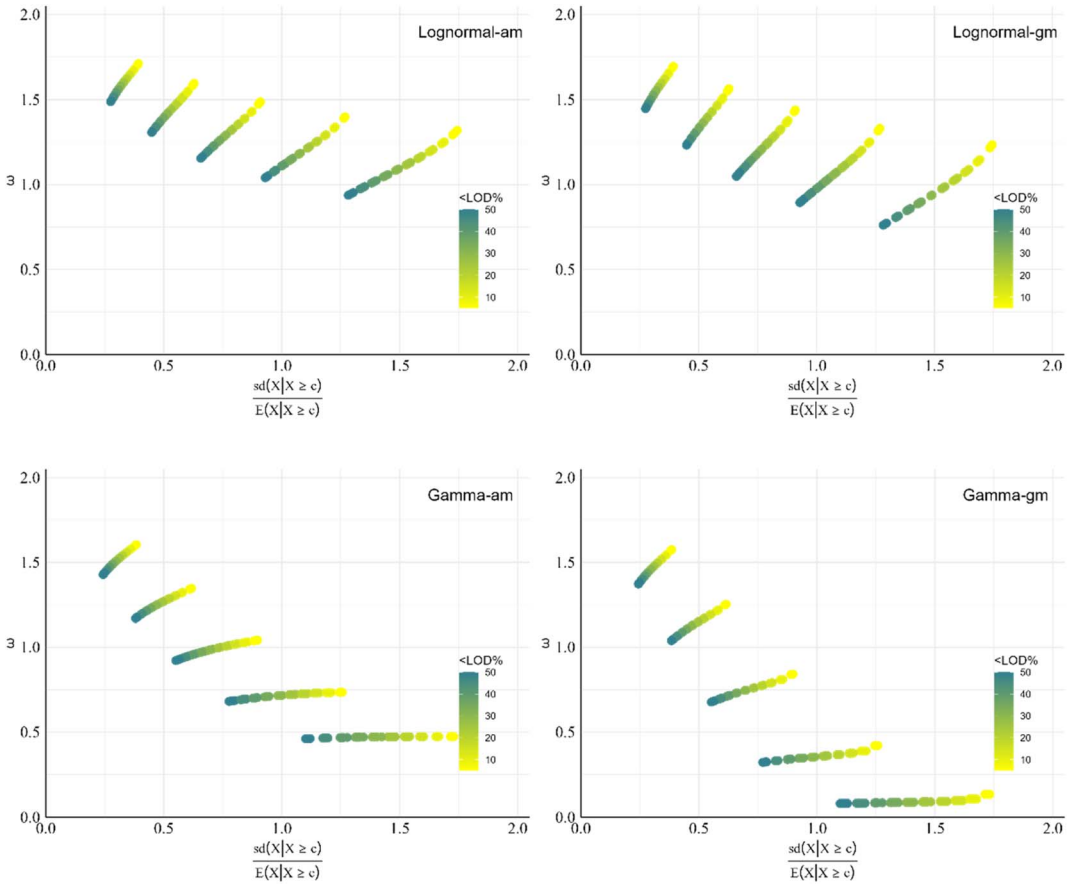

Fig. 3 Relationship between weight (ω) and $\frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$ for accurate calculation of am (left) and gm (right) in lognormal and gamma censored data.

Table 1 Models to fit the relationships of weight (ω) to LOD/2 and known characteristics of censored data $E(X|X \geq c)$, $\mathrm{sd}(X|X \geq c)$ and < LOD%)[a]

Distribution	Type of mean	Models to get ω	R^2
Lognormal	am	$\omega = a \cdot t^b$ $a = 10^{-3}\,[1482 - 9.164\,c]$ $b = -10^{-3}\,[166 + 2.569\,c]$	0.996
	gm	$\omega = a \cdot t^b$ $a = 10^{-2}[142.327 - 1.098\,c]$ $b = -10^{-3}[202.6 + 3.906\,c]$	0.992
Gamma	am	$\omega = a \cdot e^{b \cdot t}$ $a = 10^{-4}\,[1.215c^2 - 131.5\,c + 23390]$ $b = -10^{-4}[1.652c^2 + 5.432\,c + 8939]$	0.999
	gm	$\omega = a \cdot e^{b \cdot t}$ $a = 10^{-4}\,[1.403\,c^2 - 123.6\,c + 29240]$ $b = -10^{-4}\,[2.253\,c^2 + 111.1\,c + 14210]$	0.983

[a] ω is weight; $t = \frac{\mathrm{sd}(X|X \geq c)}{E(X|X \geq c)}$; c = <LOD (%) and a and b are constants.

lognormal data, this is not the case for gamma data (Table S1‡). For gamma data, the median is higher than the gm, especially for highly skewed situations (Fig S9‡). This discrepancy between gm and median indicates that gm may not effectively summarize the central tendency of gamma censored data. As recommended, reporting the am, gm and median together remains a prudent approach.

4.2 Merits of weighted LOD/2 substitution

Fig. 4 and S10‡ show the RB of weighted substitution in comparison with other sophisticated methods for the HCB and α-HCH data. Aside from the LOD/2 substitution method, the RBs of other methods are all within 15%. Notably, for the HCB data, the proposed weighted substitution method has an RB of

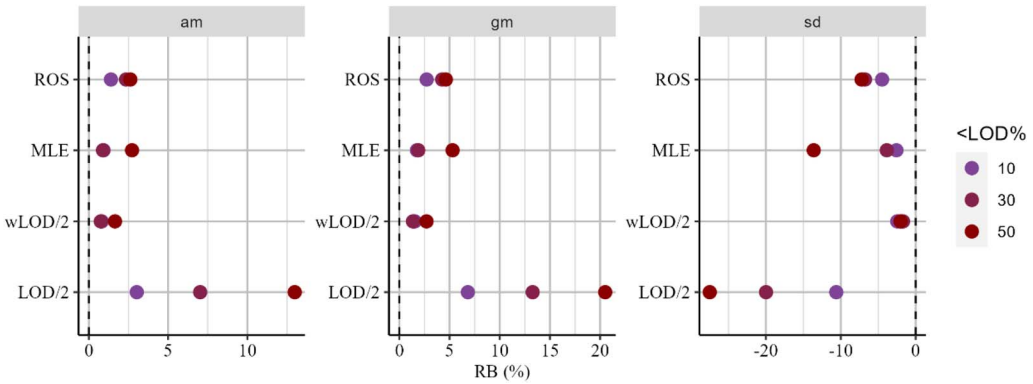

Fig. 4 Comparisons of relative bias (RB) in estimates of the am, gm and sd based on the atmospheric HCB case data. HCB is completely detected and the am, gm and sd of the whole data are known. By virtually obtaining the censored HCB data with different <LOD%, the RB could be calculated and compared among different methods. Scales are different.

around 3% and is almost unaffected by the level of censoring (Fig. 4). It is perplexing that, for the α-HCH data, the simple substitution method performs best. The weighted substitution method shows the largest RB for gm and sd when the censoring level is 50%, but it is still below 15%, which is deemed to be acceptable (Fig. S12‡). Likewise, Hewett and Ganser[19] also found that substitution methods tended to be strongly biased, but in some scenarios had the smaller error. This phenomenon may be related to the distribution of the data, as we observed

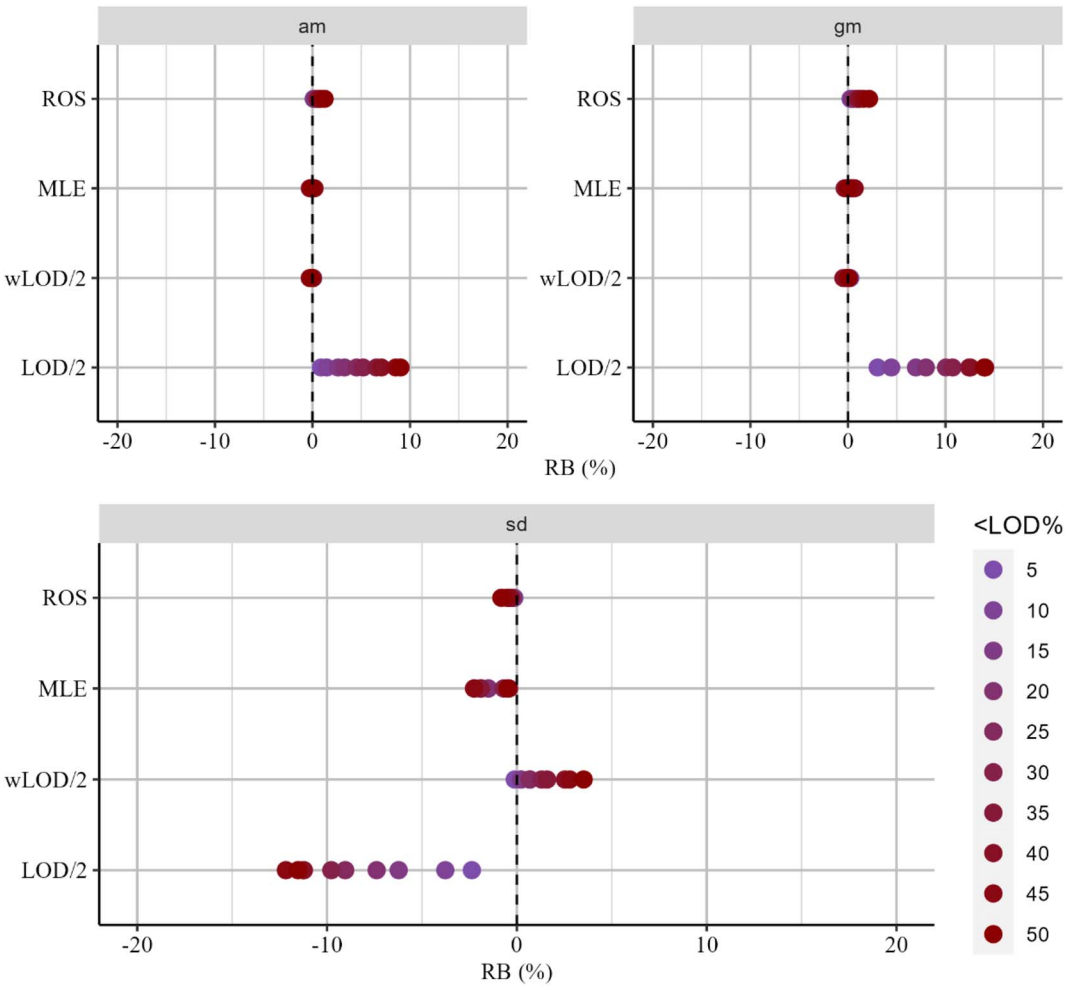

Fig. 5 Comparisons of relative bias (RB) in estimates of the am, gm and sd from the simulation study. The simulated data were generated by lognormal distribution with μ log = 1, σ log = 0.5, and $n = 50$.

a noticeable inflection in the ECDF curve of α-HCH, which makes LOD/2 the best.

Fig. 5 and 6 show the RB of the weighted substitution in comparison with other sophisticated methods in simulated data.[7,19,30] The weighted substitution yields surprisingly accurate results for mean (am and gm) estimation in both lognormal and gamma censored data (Fig. 5 and 6). For mean estimation, the weighted method is more accurate than the MLE and ROS methods. This improvement permits the choice to eliminate the bias. Regarding standard deviation (sd), its performance on lognormal data exhibits slightly more bias compared to the ROS method but is superior to the MLE method (Fig. 5). For gamma data, the bias of this method is less favorable compared to MLE (Fig. 5). Nevertheless, irrespective of the distribution type or censored level, the relative bias in sd estimation remains below 5%. This weighted method is primarily designed as an unbiased estimator for the mean and does not achieve unbiased estimation for the sd.[31] However, the method has also significantly improved the accuracy of sd estimation, which can be attributed to the more accurate mean estimation. The sample variance is given by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$. For censored values, we can write $S^2 = \frac{1}{n-1}[\sum_{i=1}^{n_M}(x_i - \overline{X})^2 + n_D \cdot (w \cdot \mathrm{LOD}/2 - \overline{X})^2]$. Both terms in the variance formula depend on the estimated mean $\bar{X}$. When $\bar{X}$ is more accurately estimated—as in our proposed method—the variance calculation more accurately reflects the true spread of the data, thereby stabilizing variance estimation. Fig. 5 and 6 exhibit the method accuracy responding to the moderately skewed distributions, as defined in a publication.[6] Fig. S11 and S12‡ show the accuracy of the weighted method with highly skewed data. The RBs of weighted substitution are within 10%, except for estimating the gm of gamma data. Overall, considering that ROS is only applicable to lognormal distributions and MLE can produce significant bias in certain situations (Fig. 6), the weighted method proposed in this paper can greatly enrich the current toolbox for accurately analyzing censored data.

Our results demonstrated that determining the distribution type of censored dataset is essential for accurate statistical estimation. For example, the precision of MLE depends on correctly specifying the likelihood function, which is derived from the probability density function of the data distribution.

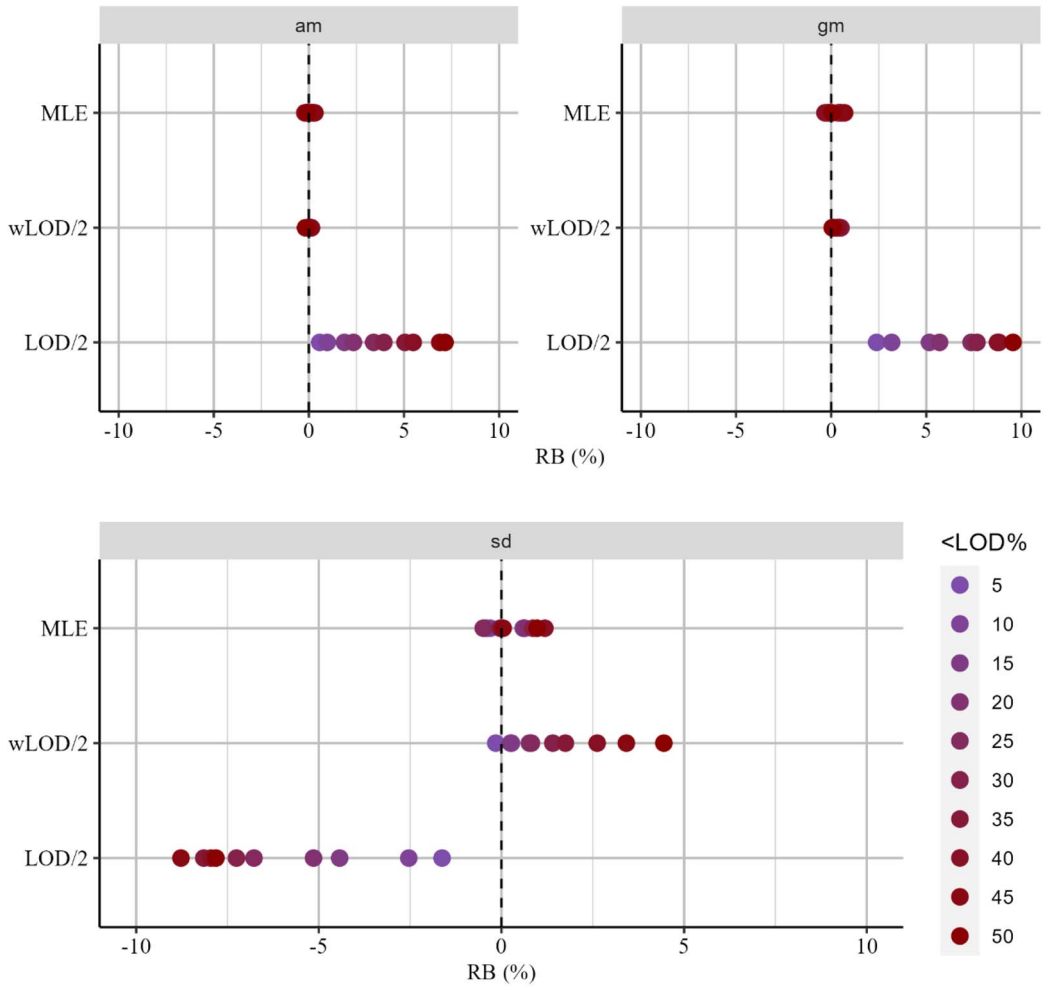

Fig. 6 Comparisons of relative bias (RB) in estimates of the am, gm and sd from the simulation study. The simulated data were generated by gamma distribution with $\alpha = 3.5$, $\beta = 1.1$, and $n = 50$.

Moreover, methods such as ROS are currently applicable only to lognormal distributions, necessitating prior knowledge of whether the data conform to this distribution. However, previous studies often assume that censored environmental data follow a lognormal distribution,[7,16,33] a supposition that this study's findings suggest that it may not always be valid. Besides, study has indicated that applying the ROS estimator directly to gamma-distributed data results in highly biased estimates and large variances.[33] In such cases, the ROS estimator proves to be not robust against deviations from the assumed distribution.

MLE, when correctly implemented using the CDF for censored data, has strong theoretical advantages, such as consistency and asymptotic efficiency. The consistency and asymptotic efficiency of MLE make it most reliable when sample sizes are sufficiently large. However, in real-world environmental applications, particularly when sample sizes are small, MLE estimates can have larger bias.[28] The weighted substitution method generally performed as well as, or better than, MLE in our simulated scenarios. Therefore, our method serves as a practical and stable alternative, particularly in cases where MLE performance is limited due to small sample sizes. We also emphasize that our proposed method is not meant to replace MLE universally, but rather to act as a flexible and computationally efficient option in specific contexts (small datasets or data scarcity). The underlying mathematical principle may explain why MLE occasionally produces larger bias in standard deviation estimates, particularly under small sample conditions. As shown in formula (1), since sd depends exponentially on both μ log and σ log, small estimation errors in these parameters can lead to larger bias in the standard deviation due to error propagation. MLE of SD for lognormal is not unbiased.

In many real-world environmental datasets, the LOD may vary across samples due to differences in analytical methods or instruments or matrix effects. In our current study, we assumed a common LOD for simplicity and clarity in presenting the method and simulations. However, handling heterogeneous LODs is important for broader applicability. The proposed ωLOD/2 method can be extended to accommodate varying LODs by applying individualized weights based on each sample's specific detection limit and its corresponding censoring level. The ωLOD/2 framework is inherently flexible and can be extended to accommodate heterogeneous LODs. Specifically, each censored value can be substituted using an individually weighted value.

For censored data, procedures for implementing distribution fitting were mentioned.[34] However, users must become proficient in the R programming language to effectively utilize these methods. Apparently, the app developed in this study incorporates methods for fitting censored data distributions, greatly enhancing the ease of use and hoping to improve the accuracy of statistical estimation.

4.3 Atmospheric OCP concentrations

Analyzing data from typical measurements is essential, particularly in environmental pollutant studies. We briefly considered the 6 atmospheric OCP datasets which contain measurements under detection limits (Fig. 1). Resorting to the convenient way to finish distribution fitting (Fig. S3‡), a solid foundation is established for selecting an appropriate likelihood function for MLE and making informed choices for ROS. Table 2 shows the estimated sample statistics for the six OCPs. We note that there is good agreement of the MLE and ωLOD/2 estimators. This proves that the weighted method provides an alternative choice to robustly calculate the mean and standard deviation of censored environmental data.

Table 2 Sample statistics of censored OCPs computed with the weighted LOD/2 substitution method

OCP	Distribution	Methods	am	sd	gm	Median
β-HCH	Gamma	ωLOD/2	0.45	0.41	0.27	NA[a]
		MLE	0.42	0.47	0.21	
γ-HCH	Gamma	ωLOD/2	2.11	1.79	1.32	1.6
		MLE	2.11	1.93	1.32	
o,*p*′-DDE	Lognormal	ωLOD/2	1.20	2.13	0.57	0.5
		MLE	1.17	2.22	0.55	
		ROS[b]	1.18	2.13	0.52	
p,*p*′-DDE	Gamma	ωLOD/2	1.97	2.40	0.85	0.9
		MLE	1.89	2.23	0.82	
o,*p*′-DDT	Gamma	ωLOD/2	3.61	3.92	1.82	1.6
		MLE	3.61	3.90	1.82	
p,*p*′-DDT	Gamma	ωLOD/2	1.19	1.18	0.62	0.65
		MLE	1.18	1.33	0.55	

[a] Median is not available because <LOD% is greater than 50%. [b] ROS is applied only for lognormal data.

Although this study focuses on atmospheric OCPs, the proposed weighted substitution approach is not limited to a specific pollutant type. As long as the data exhibit left-censoring and the distributional assumption (*e.g.*, lognormal or gamma) is reasonably appropriate, the method can be applied to a wide range of environmental contaminants, including heavy metals, PCBs, PAHs, and others, which may encounter the censored data processing challenge.[35,36]

5. Conclusions

Our weighted approach outperforms currently sophisticated methods by producing smaller errors in estimating the arithmetic mean and geometric mean for OCP data. The accuracy of standard deviation estimation is also robust with error <5% in most cases. Furthermore, we have developed a free, web-based app that integrates data fitting and statistical estimation, which is well-suited to meet the needs of environmental science and is bound to advance the field of censored data processing. The app is available at **https://lidonghuang.shinyapps.io/Censored_fitting_weighting_substitution/**. Our proposed ωLOD/2 method improves upon conventional substitution approaches (*e.g.*, LOD/2) by offering a more data-driven and adaptive solution, which may help reduce bias in estimated values, especially under high censoring. This could lead to more reliable decision-making in regulatory contexts. Agencies such as the EPA and WHO often rely on summary statistics (*e.g.*,

mean concentrations) to assess compliance, health risks, and environmental quality. However, the proposed method is currently derived based on the assumption that data follow either a lognormal or gamma distribution, which may not hold for all datasets. Therefore, future work should explore the extension of the weighting approach to other distributions, such as the Weibull distribution. Additionally, the method's performance is influenced by the accuracy of the distribution fitting process. Although functions in the EnvStats package facilitate distribution fitting, misspecification can still occur, particularly under high censoring, potentially leading to increased bias. We emphasize the importance of carefully assessing distributional assumptions, as misfitting can compromise the method's reliability. To improve robustness, we encourage practitioners to incorporate distribution fitting results from multiple sources or methods and to verify goodness-of-fit using both statistical tests and visual diagnostics.

Data availability

The data will be available upon request.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by the Research Start-up Fund of Inner Mongolia Agricultural University (No. NDYB2023-25) and Central Government's Guidance Fund for Local Scientific and Technological Development Projects (No. 2024ZY0072).

References

1 C. Yang, J. Fang, X. Sun, W. Zhang, J. Li, X. Chen, L. Yu, W. Xia, S. Xu, Z. Cai and Y. Li, Prenatal exposure to organochlorine pesticides and infant growth: A longitudinal study, *Environ. Int.*, 2021, **148**, 106374, DOI: 10.1016/j.envint.2020.106374.

2 S.-H. Seo, S.-D. Choi, S. Batterman and Y.-S. Chang, Health risk assessment of exposure to organochlorine pesticides in the general population in Seoul, Korea over 12 years: A cross-sectional epidemiological study, *J. Hazard. Mater.*, 2022, **424**, 127381, DOI: 10.1016/j.jhazmat.2021.127381.

3 X. Wu, A. Chen, Z. Yuan, H. Kang and Z. Xie, Atmospheric organochlorine pesticides (OCPs) and polychlorinated biphenyls (PCBs) in the Antarctic marginal seas: distribution, sources and transportation, *Chemosphere*, 2020, **258**, 127359.

4 C. Wang, X. Wang, X. Yuan, J. Ren and P. Gong, Organochlorine pesticides and polychlorinated biphenyls in air, grass and yak butter from Namco in the central Tibetan Plateau, *Environ. Pollut.*, 2015, **201**, 50–57.

5 J. Ren, X. Wang, C. Wang, P. Gong and T. Yao, Atmospheric processes of organic pollutants over a remote lake on the central Tibetan Plateau: implications for regional cycling, *Atmos. Chem. Phys.*, 2017, **17**(2), 1401–1415.

6 B. J. George, L. Gains-Germain, K. Broms, K. Black, M. Furman, M. D. Hays, K. W. Thomas and J. E. Simmons, Censoring Trace-Level Environmental Data: Statistical Analysis Considerations to Limit Bias, *Environ. Sci. Technol.*, 2021, **55**(6), 3786–3795, DOI: 10.1021/acs.est.0c02256.

7 R. A. Hites, Correcting for Censored Environmental Measurements, *Environ. Sci. Technol.*, 2019, **53**(19), 11059–11060, DOI: 10.1021/acs.est.9b05042.

8 L. Huang, K. Bradshaw, J. Grosskleg and S. D. Siciliano, Assessing Space, Time, and Remediation Contribution to Soil Pollutant Variation near the Detection Limit Using Hurdle Models to Account for a Large Proportion of Nondetectable Results, *Environ. Sci. Technol.*, 2019, **53**(12), 6824–6833, DOI: 10.1021/acs.est.8b07110.

9 C. P. Haslauer, J. R. Meyer, A. Bárdossy and B. L. Parker, Estimating a Representative Value and Proportion of True Zeros for Censored Analytical Data with Applications to Contaminated Site Assessment, *Environ. Sci. Technol.*, 2017, **51**(13), 7502–7510, DOI: 10.1021/acs.est.6b05385.

10 L. Fu and Y.-G. Wang, Nonparametric Rank Regression for Analyzing Water Quality Concentration Data with Multiple Detection Limits, *Environ. Sci. Technol.*, 2011, **45**(4), 1481–1489, DOI: 10.1021/es101304h.

11 R. A. Hites, A Statistical Approach for Left-Censored Data: Distributions of Atmospheric Polychlorinated Biphenyl Concentrations near the Great Lakes as a Case Study, *Environ. Sci. Technol. Lett.*, 2015, **2**(9), 250–254, DOI: 10.1021/acs.estlett.5b00223.

12 T. Amemiya, Tobit models: a survey, *J. Econom.*, 1984, **24**(1), 3–61, DOI: 10.1016/0304-4076(84)90074-5.

13 A. Chesher, D. Kim and A. M. Rosen, IV methods for Tobit models, *J. Econom.*, 2023, **235**(2), 1700–1724, DOI: 10.1016/j.jeconom.2023.01.010.

14 R. T. Carson and Y. Sun, The Tobit model with a non-zero threshold, *J. Econom.*, 2007, **10**(3), 488–502, DOI: 10.1111/j.1368-423X.2007.00218.x.

15 E. Brankov, S. T. Rao and P. S. Porter, Identifying Pollution Source Regions Using Multiply Censored Data, *Environ. Sci. Technol.*, 1999, **33**(13), 2273–2277, DOI: 10.1021/es980479j.

16 H. G. Mikkonen, B. O. Clarke, R. Dasika, C. J. Wallis and S. M. Reichman, Evaluation of methods for managing censored results when calculating the geometric mean, *Chemosphere*, 2018, **191**, 412–416, DOI: 10.1016/j.chemosphere.2017.10.038.

17 A. J. Turkson, F. Ayiah-Mensah and V. Nimoh, Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review, *Int. J. Math. Math. Sci.*, 2021, **2021**(1), 9307475, DOI: 10.1155/2021/9307475.

18 N. E. Breslow, Analysis of Survival Data under the Proportional Hazards Model, *Int. Stat. Rev.*, 1975, **43**(1), 45–57, DOI: 10.2307/1402659.

19 P. Hewett and G. H. Ganser, A comparison of several methods for analyzing censored data, *Ann. Occup. Hyg.*, 2007, **51**(7), 611–632, DOI: **10.1093/annhyg/mem045**.

20 D. R. Helsel, Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it, *Chemosphere*, 2006, **65**(11), 2434–2439, DOI: **10.1016/j.chemosphere.2006.04.051**.

21 B. J. George, K. W. Thomas and J. E. Simmons, Response to Comment on "Censoring Trace-Level Environmental Data: Statistical Analysis Considerations to Limit Bias", *Environ. Sci. Technol.*, 2021, **55**(22), 15556–15557, DOI: **10.1021/acs.est.1c06431**.

22 CDC, *Fourth National Report on Human Exposure to Environmental Chemicals*, 2021.

23 S. P. Millard, *EnvStats: an R Package for Environmental Statistics*: Springer. 2013.

24 M. Furman, K. W. Thomas and B. J. George, Separating Measurement Error and Signal in Environmental Data: Use of Replicates to Address Uncertainty, *Environ. Sci. Technol.*, 2023, **57**(41), 15356–15365, DOI: **10.1021/acs.est.3c02231**.

25 E. Newton and R. Rudel, Estimating Correlation with Multiply Censored Data Arising from the Adjustment of Singly Censored Data, *Environ. Sci. Technol.*, 2007, **41**(1), 221–228, DOI: **10.1021/es0608444**.

26 Y. Jin, M. J. Hein, J. A. Deddens and C. J. Hines, Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS, *Ann. Occup. Hyg.*, 2011, **55**(1), 97–112.

27 L. W. Stanek, N. Grokhowsky, B. J. George and K. W. Thomas, Assessing lead exposure in U.S. pregnant women using biological and residential measurements, *Sci. Total Environ.*, 2023, **905**, 167135, DOI: **10.1016/j.scitotenv.2023.167135**.

28 M. A. Tekindal, B. D. Erdoğan and Y. Yavuz, Evaluating Left-Censored Data Through Substitution, Parametric, Semi-parametric, and Nonparametric Methods: A Simulation Study, *Interdiscip. Sci.:Comput. Life Sci.*, 2017, **9**(2), 153–172, DOI: **10.1007/s12539-015-0132-9**.

29 R-Core-Team, *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024, **https://www.R-project.org/**. 2024.

30 B. W. Gillespie, Q. Chen, H. Reichert, A. Franzblau, E. Hedgeman, J. Lepkowski, P. Adriaens, A. Demond, W. Luksemburg and D. H. Garabrant, Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator, *Epidemiology*, 2010, **21**(4), S64–S70.

31 A. H. El-Shaarawi and S. R. Esterby, Replacement of censored observations by a constant: An evaluation, *Water Res.*, 1992, **26**(6), 835–844, DOI: **10.1016/0043-1354(92)90015-V**.

32 N. Shoari, J.-S. Dubé and S. e. Chenouri, Estimating the mean and standard deviation of environmental data with below detection limit observations: considering highly skewed data and model misspecification, *Chemosphere*, 2015, **138**, 599–608, DOI: **10.1016/j.chemosphere.2015.07.009**.

33 R. H. Shumway, R. S. Azari and M. Kayhanian, Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits, *Environ. Sci. Technol.*, 2002, **36**(15), 3345–3353, DOI: **10.1021/es0111129**.

34 D. R. Helsel, Computing Summary Statistics and Totals, in *Statistics for Censored Environmental Data Using Minitab® and R*, 2011, pp. 62–98.

35 A. N. Fendrich, E. Van Eynde, D. M. Stasinopoulos, R. A. Rigby, F. Y. Mezquita and P. Panagos, Modeling arsenic in European topsoils with a coupled semiparametric (GAMLSS-RF) model for censored data, *Environ. Int.*, 2024, **185**, 108544, DOI: **10.1016/j.envint.2024.108544**.

36 E. F. Eastoe, C. J. Halsall, J. E. Heffernan and H. Hung, A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case study from the Canadian Arctic, *Atmos. Environ.*, 2006, **40**(34), 6528–6540, DOI: **10.1016/j.atmosenv.2006.05.073**.

