Small Values in Big Data: The Continuing Need for Appropriate Metadata

Craig A. Stow[1], Katherine E. Webster[2], Tyler Wagner[3], Noah Lottig[4], Patricia A. Soranno[2], YoonKyung Cha[5]

[1]National Oceanic and Atmospheric Administration Great Lakes Environmental Research Laboratory, Ann Arbor, MI 48176 USA

[2]Michigan State University, Dept. Fisheries & Wildlife, East Lansing, MI 48824 USA

[3]U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Unit, The Pennsylvania State University, 402 Forest Resources Building, University Park, PA, 16802

[4]Univ Wisconsin, Center for Limnology, Boulder Jct, WI USA

[5]Univ Seoul, School Environmental Engineering, Seoul, South Korea

Abstract

Compiling data from disparate sources to address pressing ecological issues is increasingly common. Many ecological datasets contain left-censored data – observations below an analytical detection limit. Studies from single and typically small datasets show that common approaches for handling censored data — e.g., deletion or substituting fixed values — result in systematic biases. However, no studies have explored the degree to which the documentation and presence of censored data influence outcomes from large, multi-sourced datasets. We describe left-censored data in a lake water quality database assembled from 74 sources and illustrate the challenges of dealing with small values in big data, including detection limits that are absent, range widely, and show trends over time. We show that substitutions of censored data can also bias analyses using 'big data' datasets, that censored data can be effectively handled with modern quantitative approaches, but that such approaches rely on accurate metadata that describe treatment of censored data from each source.

Introduction

Data sharing is an increasing expectation in the sciences[1-3]. This outlook arises from the recognition that data are expensive and should be made widely available for maximum utility, as well as the view that information funded by taxpayers should be accessible. Although there have been concerns that users of such data are simply “datavores” or perhaps worse, “research parasites” [4], there are many scientific gains to be made from assembling data from diverse sources and harmonizing them into a consistent format for further research. The environmental sciences, in particular, stand to benefit as we investigate phenomena occurring across broad spatial and temporal scales[5-7].

Comprehensive metadata are essential to interpret large, integrated databases so that data provenance and context are retained[1,8], and to reduce the chance that patterns accidentally arise as artifacts of differing observational protocols. Complete metadata should accurately describe the “censored” observations, which result when measured samples have values that are either too high or low to be quantified (supplemental box). Samples that are below a lower detection limit are most common and are termed “left-censored”. Examples include nutrient and chemical concentrations that fall below the detection limit of the analytical approach[9-10]. Though less common, “right-censoring” may also occur when, for example, concentrated aqueous samples are not adequately diluted before analysis or when Secchi depth, a measure of water clarity, exceeds the lake depth[11].

Analyzing data containing censored observations may be complicated by the fact that detection limits for the same characteristic can differ depending on the measurement protocols used, and may change over time. Ideally, metadata in a harmonized database would indicate which observations are censored and the detection limit for each censored observation. However, even basic metadata can be lacking in data repositories containing data from many sources[8]. Thus, it is important to consider whether the censored observations are sufficiently well-documented in ecological datasets to rigorously use them in analyses of compiled datasets.

Two common approaches for treating left-censored data include: 1) discarding the censored observations or 2) substituting a value including: the detection limit, half the detection limit, or zero. Under limited circumstances, these informal approaches may not strongly influence the conclusions derived from the data analysis. For example, qualitative pattern assessment may not be affected, particularly if the proportion of censored observations is low, and their range is small relative to the overall data range. However, censored data contain information, which will be improperly represented when observations are discarded or substitution is used, possibly influencing inference, particularly when they comprise higher proportions of the database. Additionally, even if the overall proportion of censored observations is small, censoring may be disproportionately high in some groups within the data, causing misleading comparisons.

Rigorous approaches to accommodate censored data have long been available[12-14]. Helsel [15-18], Antweiler and Taylor[19], and Antweiler[20] stressed the challenges of analyzing censored data and presented methods to analyze datasets containing censored observations. However, these approaches still require accurate censoring metadata for all observations.

Our goal was to examine censored data properties in commonly-measured ecological variables that have been harmonized into a large, integrated database to determine the effect of censored data on ecological inference. Because such integrated databases are becoming increasingly common, the potential biases due to censored data invites investigation. We used a large, harmonized water quality database compiled from 76 sources[21-22]. Our objectives were to quantify: a) the proportion of datasets and data values with sufficient metadata to confidently identify censored observations; b) variation in reported detection limits across sources and through time in the last several decades of water quality sampling; and c) the effect of three strategies for dealing with censored observations on a simple water quality model and whether the proportion of censored observations influences that effect. Our results highlight the need for accurate documentation and metadata.

Methods

We draw on our experience in developing LAGOS-NE (**LA**ke multi-scaled **G**e**OS**patial & temporal database – Northeast and Midwest lakes), a lake water quality database with data from 17 northeastern USA states[21]. LAGOS-NE version 1.087.1 includes contributions from 76 state, federal, tribal, university, citizen science, and non-profit monitoring programs with chlorophyll a, total nitrogen, and total phosphorus (CHLa, TN, and TP, respectively) measurements in lake surface waters. Data from two monitoring programs, consisting of 1 and 5 total observations, were omitted prior to our analysis. The number of observations and programs supplying data for each variable ranged, respectively, from 40,670 to 209,732 and from 33 to 66 (Table 1); most data were collected between 1970 and 2013.

During the creation of LAGOS-NE, codes that documented censor status and whether or not the source program provided detection limits were assigned to each observation. Data providers indicated values were censored in multiple ways: (a) explicit detection limits (DL) were provided with each value; (b) DLs were assumed to be the reported value when tags such as '<' were provided; and (c) DLs were provided in the metadata but not specified in the dataset. Based on these codes, we summarized the number of programs and corresponding number of observations that had DL information and the proportion of LAGOS-NE data that was comprised of censored observations for each water quality variable. We used, respectively, statistical summaries and cumulative frequency distributions compiled at decadal time steps to provide insights into variation in DLs among programs and over time.

Prior to finalizing LAGOS-NE, we deleted a small number of non-censored that values were reported as zero (351, 40 and 266 for CHLa, TN and TP, respectively). We made the decision to delete these, because it was unclear if these values were true zeroes, rounding artifacts, or substituted values and because bivariate plots with related variables indicated, in many cases, that these were outlier values.

To demonstrate the effect that data censoring can have on quantitative analyses we simulated a large dataset with known censoring patterns. The simulated data represent a log-linear relationship between TP and CHLa concentrations using parameter values previously estimated from a subset of LAGOS-NE lakes[23]. We performed simulations where the proportion of censoring was set to 5, 15, and 30% of the simulated data. For each of the three sets of simulations, we generated 100 datasets consisting of 10,000 lakes each. The intercept, slope and residual standard deviation used to generate the data were -0.24, 0.83, and 0.40, respectively. For each simulated dataset, the response variable, CHLa, was left-censored at 5, 15, or 30%. We then analyzed each dataset using linear regression where the censored values were estimated iteratively and constrained to fall below the detection limit[24-25], and three naïve approaches where: (1) censored values were omitted, (2) censored values were set to the detection limit, and (3) censored values were set to half the detection limit. All models were fitted using Bayesian estimation. Diffuse normal priors (N[0,1000]) were used for the intercept and slope parameters and a diffuse uniform prior (Unif[0,10]) was used for the residual standard deviation using JAGS in the R2jags package[26], run from within R version 3.3.0[27]. We ran three parallel Markov chains beginning each chain with different values. From a total of 10,000 samples from the posterior distribution the first 5,000 samples of each chain were discarded for a total of 15,000 samples used to characterize the posterior distributions. We assessed convergence for all parameters both visually (trace plots), as well as with the Brooks-Gelman-Rubin statistic. During each simulation the estimated values of the intercept, slope, and residual standard deviation were compared to the true values used in the data generating process to calculate the resultant biases.

Results

Depending on the water quality variable, 39.4 to 60.6 % of programs documented censored observations either within the database or in accompanying metadata (Table 1a). Despite substantial proportion of programs that did not provide DL information, their contributions constituted less than 20 % of the observations in LAGOS-NE, suggesting that larger lake monitoring programs typically had more information on censored data. Further, censored observations comprised a small percentage of the database, 2.4 % or less for all three water quality variables (Table 1b).

The wide range of ways that censored data were identified in the original program datasets complicated harmonization. For example, observations could be associated with specific DLs, DLs could be documented program-wide, or DLs could be identified as tagged values or even, in one case, inserted as negative numbers in the database. The percentage of observations with specified DLs differed depending on the water quality variable. For CHLa, TN, and TP, respectively, 23, 66 and 28 % of observations had the DL specified for each observation; 19, 2, and 42 % of observations had DLs assigned through metadata or as tags; and the remaining 38, 18, and 14 % of observations were from datasets with a mixture of censoring strategies. A few of the latter programs provided databases with data collected over multiple decades and may have changed specification of censored data within their database over time.

The extent to which individual programs substituted values when concentrations were less than the DL cannot be fully evaluated. For censored observations that had associated DLs specified, respectively, 7.5, 0, and 12.7 % of observations were equal to one-half the DL and 42.1, 1.6, and 16.0 % observations were equal to the DL for CHLa, TN, and TP. Some programs reported non-censored observations with concentrations less than the reported DL, possibly indicating that the reported DL was an overall method DL, not batch-specific. This disparity of reporting approaches for censored observations was one of the most challenging aspects of data harmonization.

Further complexity for data users of LAGOS-NE was the wide range of DLs (Table 1b). Reported detection limits differed by over two orders of magnitude for CHLa and TP (Table 1b); six DLs for TP were very high and exceeded 100 µg/L, with a maximum at 570. Despite large ranges, however, median DLs were low, respectively, 1, 50 and 2 µg/L for CHLa, TN and TP.

Finally, we compared the overall distribution of DLs with those for data collected prior to 2000 and in the 2000 and 2010 decades (Figure 1). Temporal patterns in detection limits differed among the three water chemistry variables. DLs for CHLa were most consistent over the three time periods, with a only a small percentage having DLs exceeding 1 ug/L. In contrast, DLs for TN and TP differed in cumulative frequency over time. For TN, DLs for samples collected prior to 2000 included both lower and higher values compared to other time periods and overall (Figure 1a). For TP, data collected prior to 2000 had lower DLs compared to later years with 70% of DL values less than 10 µg/L. The time period prior to 2000 did have a higher frequency of DLs equal to and greater than 20 µg/L compared to later years, including half of the six DL's over 100 and the two values exceeding 200. In subsequent decades, the DL for TP analyses shifted towards a dominance of DL equal to 10 µg/L. These patterns suggest, at least for TP, that while maximum detection limits have declined over time, the majority of earlier data was analyzed under protocols with generally lower DLs. We speculate that this might be due to increased automation in laboratories combined with a tradeoff of sacrificing lower sensitivity at lower ends of the concentration range. The results provide cautions that systematic differences in DL within the database have the potential to generate artifacts that interfere with trends and patterns in the data, particularly influencing analyses based on low concentrations.

Our simulation study of the effects of different replacement strategies for censored data on parameter estimation provide further evidence for careful consideration of how censored observations are treated in large datasets. Regression lines generated from one of the 100 simulated data sets of 10,000 lakes help visualize the problem that occurs using various methods to accommodate the censored observations (Figure 2a). In this specific result, the "true" regression and censored model lines are essentially coincident, indicating that the censored model closely replicates the truth. The lines generated by omitting the censored observations and setting the censored observations to the detection limit are similar to one-another, both with intercepts that are higher and slopes that are lower than those of the "true" model. In contrast, the line that results from setting the censored observations to half the detection limit has an intercept that is lower and a slope that is higher than the true model.

This specific result is indicative of the general pattern that becomes apparent from the 100 simulations (Figure 2b). Omitting censored observations or setting them to the detection limit causes negatively biased slopes, positively biased intercepts, and negatively biased standard deviations. However, when the censored observations are set to half the detection limit, the slope, intercept, and standard deviation biases are reversed. For all three methods the size of the bias increases with the proportion of censored observations. Concurrently, the censored model remains unbiased, even when 30% of the observations were censored.

Discussion

We offer a cautionary tale regarding potential problems posed by censored data, for which approaches to address them have been documented in the literature for many years. However, adding to the analytical issues raised in the past, the censored data in LAGOS-NE v1.087.1 are likely characteristic of other large, harmonized, environmental databases and illustrate that despite a history of documentation, problems persist, and new uncertainties introduced due to differences in analytical procedures and data reporting among monitoring programs. While the proportion of values clearly identifiable as below detection was small, there remained a proportion of observations showing symptoms consistent with having been substituted, as well as a small number that we labeled as "missing" because it was unclear if they were truly zero or if their missingness was a detection limit artifact. This inability to clearly differentiate censored observations puts users of compiled data in a difficult position; we discarded a small number of observations for lack of a clearly superior alternative, given the limitations of the supporting metadata.

Our results highlight the need for standard reporting of censored data for these common water quality variables and identify complexities inherent in combining data from disparate sources. Additionally, our results support findings of Sprague et al.[8] regarding difficulties in combining datasets. In the case of LAGOS-NE, many of the limitations described in Sprague et al.[8] were minimized because we solicited data directly from the program maintainers and requested metadata information regarding aspects such as units, methods, chemical species and detection limits and associated data tags[20]. In fact, if the dataset did not contain sufficient metadata we did not consider it for inclusion in LAGOS-NE; however even with substantial metadata, censored observation documentation was sometimes ambiguous.

Further, our simulation study showed how handling of censored data could influence common analyses, such as regression modeling. The approach we have demonstrated is useful for linear regression modeling; other approaches are available for different applications. For example, the Bayesian hurdle model can use one set of predictor variables to predict which response variable observations are below detection, and another set to estimate the value of the response variable for those observations above the detection limit[28]. An important outcome of our analysis shows that such biases do not diminish with sample size. Thus, if quantified estimates are needed, as they are for most statistical analyses of large datasets, then choosing methods to appropriately incorporate the censored observations is necessary, and metadata documentation of censoring is critical.

Harmonizing datasets from multiple sources offers great benefits, but also presents challenges, many of which can be overcome with accurate metadata documenting the nuances of the assembled data. The first major challenge that we documented is the wide range of strategies for documenting DLs and censored observations among data sources. This challenge makes data harmonization especially time-consuming. The second major challenge more for users of the database is the changes in reported DL from the 1970's to present, the period when many ecological datasets have been collected. These changes could bias trend detection in lower concentrations of ecological variables such as nutrients. Although problems posed by improper censored handling data are well-documented, and approaches to accommodate censored observations are available when censored status is fully known, we find that the problem persists. The temptation to treat left-censored values cavalierly may arise because, for many environmental applications, low values indicate the absence of contamination, and thus are of minimal concern. However, using substitution or discarding low values resulted in biased estimation even when the proportion of censored values was small and the number of observations was large. Our regression analysis example demonstrates that contemporary computational approaches make rigorous treatment of censored observations straightforward, if the metadata include adequate documentation. For censored data this documentation should include a clear indication of which observations were censored and a specification of the detection limit for each censored observation. Thorough compilation of detailed metadata in the database harmonization process and attention to metadata during statistical analyses by the user remain critical for successful research efforts relying on big data.

Acknowledgements

This study was supported by the U.S. National Science Foundation MacroSystems Biology Program in the Biological Sciences Directorate (EF-1065786, EF-1065649, and EF-1065818) to all authors; and the USDA National Institute of Food and Agriculture, Hatch Project no. 176820 to PAS. We thank the members of the Cross-scale Interactions Research Project who created LAGOS-NE, as well as the many individuals who sampled the thousands of lakes, analyzed the hundreds of thousands of samples, documented and maintained their datasets, and then shared them with us. GLERL contribution number 1878. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The authors declare no competing financial interest.

References

(1) Soranno, P. A.; Cheruvelil, K. S.; Elliott, K. C.; Montgomery, G. M., It's Good to Share: Why Environmental Scientists' Ethics Are Out of Date. *Bioscience* **2015,** *65*, (1), 69-73.

(2) McNutt, M.; Lehnert, K.; Hanson, B.; Nosek, B. A.; Ellison, A. M.; King, J. L., Liberating field science samples and data. *Science* **2016,** *351*, (6277), 1024-1026.

(3) Schimel, D., Open data. *Frontiers in Ecology and the Environment* **2017**, 15, (4), 175.

(4) McNutt, M., # IAmAResearchParasite. *Science* **2016,** *351*, (6277), 1005-1005.

(5) Heffernan, J. B.; Soranno, P. A.; Angilletta, M. J.; Buckley, L. B.; Gruner, D. S.; Keitt, T. H.; Kellner, J. R.; Kominoski, J. S.; Rocha, A. V.; Xiao, J. F.; Harms, T. K.; Goring, S. J.; Koenig, L. E.; McDowell, W. H.; Powell, H.; Richardson, A. D.; Stow, C. A.; Vargas, R.; Weathers, K. C., Macrosystems ecology: understanding ecological patterns and processes at continental scales. *Frontiers in Ecology and the Environment* **2014,** *12*, (1), 5-14.

(6) O'Reilly, C. M.; Sharma, S.; Gray, D. K.; Hampton, S. E.; Read, J. S.; Rowley, R. J.; Schneider, P.; Lenters, J. D.; McIntyre, P. B.; Kraemer, B. M.; Weyhenmeyer, G. A.; Straile, D.; Dong, B.; Adrian, R.; Allan, M. G.; Anneville, O.; Arvola, L.; Austin, J.; Bailey, J. L.; Baron, J. S.; Brookes, J. D.; de Eyto, E.; Dokulil, M. T.; Hamilton, D. P.; Havens, K.; Hetherington, A. L.; Higgins, S. N.; Hook, S.; Izmest'eva, L. R.; Joehnk, K. D.; Kangur, K.; Kasprzak, P.; Kumagai, M.; Kuusisto, E.; Leshkevich, G.; Livingstone, D. M.; MacIntyre, S.; May, L.; Melack, J. M.; Mueller-Navarra, D. C.; Naumenko, M.; Noges, P.; Noges, T.; North, R. P.; Plisnier, P. D.; Rigosi, A.; Rimmer, A.; Rogora, M.; Rudstam, L. G.; Rusak, J. A.; Salmaso, N.; Samal, N. R.; Schindler, D. E.; Schladow, S. G.; Schmid, M.; Schmidt, S. R.; Silow, E.; Soylu, M. E.; Teubner, K.; Verburg, P.; Voutilainen, A.; Watkinson, A.;

Williamson, C. E.; Zhang, G. Q., Rapid and highly variable warming of lake surface waters around the globe. *Geophys Res Lett* **2015,** *42*, (24), 10773-10781.

(7) LaDeau, S. L.; Han, B. A.; Rosi-Marshall, E. J.; Weathers, K. C., The Next Decade of Big Data in Ecosystem Science. *Ecosystems* **2017,** *20*, (2), 274-283.

(8) Sprague, L. A.; Oelsner, G. P.; Argue, D. M., Challenges with secondary use of multi-source water-quality data in the United States. *Water Res* **2017,** *110*, 252-261.

(9) Alexander, R. B.; Smith, R. A., Trends in the nutrient enrichment of US rivers during the late 20th century and their relation to changes in probable stream trophic conditions. *Limnol Oceanogr* **2006,** *51*, (1), 639-654.

(10) Phillips, P. J.; Schubert, C.; Argue, D.; Fisher, I.; Furlong, E. T.; Foreman, W.; Gray, J.; Chalmers, A., Concentrations of hormones, pharmaceuticals and other micropollutants in groundwater affected by septic systems in New England and New York. *Sci Total Environ* **2015,** *512*, 43-54.

(11) Carstensen, J., Censored data regression: Statistical methods for analyzing Secchi transparency in shallow systems. *Limnol Oceanogr-Meth* **2010,** *8*, 376-385.

(12) Gilliom, R. J.; Helsel, D. R., Estimation of Distributional Parameters for Censored Trace Level Water-Quality Data .1. Estimation Techniques. *Water Resour Res* **1986,** *22*, (2), 135-146.

(13) Helsel, D. R.; Gilliom, R. J., Estimation of Distributional Parameters for Censored Trace Level Water-Quality Data .2. Verification and Applications. *Water Resour Res* **1986,** *22*, (2), 147-155.

(14) Elshaarawi, A. H.; Dolan, D. M., Maximum-Likelihood Estimation of Water-Quality Concentrations from Censored-Data. *Can J Fish Aquat Sci* **1989,** *46*, (6), 1033-1039.

(15) Helsel, D. R., More than obvious: Better methods for interpreting nondetect data. *Environ Sci Technol* **2005,** *39*, (20), 419a-423a.

(16) Helsel, D. R., Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **2006,** *65*, (11), 2434-2439.

(17) Helsel, D., Much Ado About Next to Nothing: Incorporating Nondetects in Science. *Ann Occup Hyg* **2010,** *54*, (3), 257-262.

(18) Helsel DR. 2012. Statistics for Censored Environmental Data using Minitab and R. John Wiley & Sons, Inc. Hoboken, NJ.

(19) Antweiler, R. C.; Taylor, H. E., Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environ Sci Technol* **2008,** *42*, (10), 3732-3738.

(20) Antweiler, R. C., Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets. II. Group Comparisons. *Environ Sci Technol* **2015,** *49*, (22), 13439-13446.

(21) Soranno, P. A.; Bissell, E. G.; Cheruvelil, K. S.; Christel, S. T.; Collins, S. M.; Fergus, C. E.; Filstrup, C. T.; Lapierre, J. F.; Lottig, N. R.; Oliver, S. K.; Scott, C. E.; Smith, N. J.; Stopyak, S.; Yuan, S.; Bremigan, M. T.; Downing, J. A.; Gries, C.; Henry, E. N.; Skaff, N. K.; Stanley, E. H.; Stow, C. A.; Tan, P. N.; Wagner, T.; Webster, K. E., Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *Gigascience* **2015,** *4,* (1), 1-15.

(22) Soranno, P.A., L.C. Bacon, M. Beauchene, K.E. Bednar, E.G. Bissell, C.K. Boudreau, M.G. Boyer, M.T. Bremigan, S.R. Carpenter, J.W. Carr and 70 co-authors. LAGOS-NE: A multi-

scaled geospatial temporal database of lake ecological context and water quality for thousands of U.S. Lakes. *GigaScience*. **2017,** *12,* (12), 1-22.

(23) Wagner, T.; Soranno, P. A.; Webster, K. E.; Cheruvelil, K. S., Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes. *Freshwater Biol* **2011,** *56*, (9), 1811-1824.

(24) Gelman A, and Hill J. 2007. Data Analysis Using Regression and Multilvel/Hierarchical Models. Cambridge University Press.

(25) Yun, J.; Qian, S. S., A Hierarchical Model for Estimating Long-Term Trend of Atrazine Concentration in the Surface Water of the Contiguous US. *J Am Water Resour As* **2015,** *51*, (4), 1128-1137.

(26) Su, Y.-S., and Yajima, M. 2015. R2jags: Using r to run 'jags'. R package version 0.5-7: https://CRAN.R-project.org/package=R2jags.

(27) R Core Team. R: A Language and Environment for Statistical Computing. 2016. Vienna Austria. https://www.R-project.org/

(28) Cha, Y.; Park, S. S.; Kim, K.; Byeon, M.; Stow, C. A., Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. *Water Resour Res* **2014,** *50*, (3), 2518-2532.

Table Captions

Table 1. Overview of censored and non-censored data in the LAGOS-NE database for each water quality variable. (a) The number and percentages of individual programs supplying datasets with and without DL information and the corresponding number and percentage of observations. (b) The number of censored observations within LAGOS-NE and summary statistics of DL for censored values.

Table 1

		Water quality variable		
Measure		**CHLa**	**TN**	**TP**
(a) Programs with and without DL information				
Number of programs	*n*	*58*	*33*	*66*
Percent with DL information	%	43.1	39.4	60.6
Percent with no DL information	%	56.9	60.6	39.4
Number of observations	*n*	*209732*	*41670*	*158968*
Percent from programs with DL	%	80.6	85.6	83.1
Percent from programs with no DL	%	19.4	14.4	16.9
(b) DL from censored observations				
Number of censored observations	*n*	*5088*	*192*	*3264*
	% of total	2.43	0.46	2.05
Concentration (µg/L)	median	1	84	10
	mean	0.99	145.3	9.0
	min	0.03	20	0.3
	max	10	280	570

Figure Captions

Figure 1

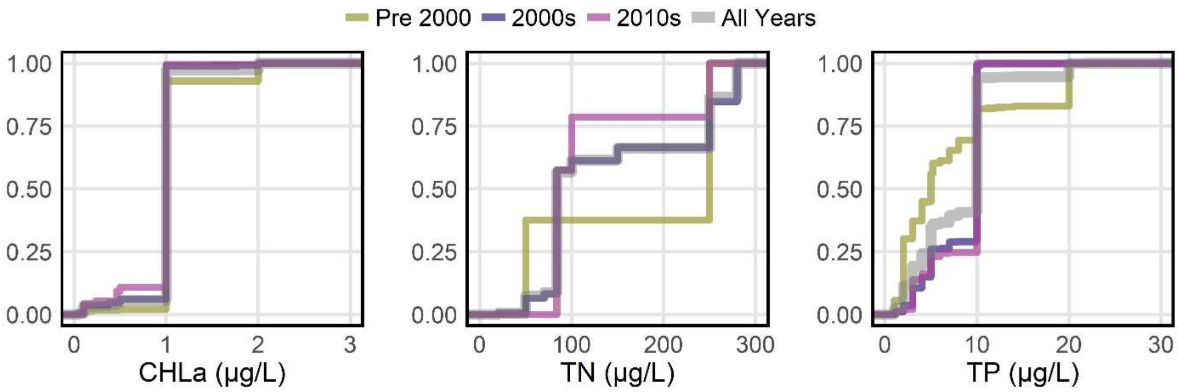
Cumulative frequency distribution plots of detection limits for censored observations in LAGOS-NE. Distributions of all DLs and those within decadal time intervals are shown. The x-axis for TP and CHLa plots, respectively, were truncated to 30 and 3 µg/L to better capture the majority of observations, thus eliminating 84 and 18 observations. Summary statistics are in Table 1.

Figure 2

(a) One realization from a simulation representing the log-linear relationship between total phosphorus (predictor variable) and chlorophyll *a* (response variable) in north temperate lakes. Dots represent values from individual lakes (n = 10,000) and open dots represent censored observations, where 30% of the observations are left-censored. Solid lines are posterior mean regression lines from a censored regression model and three naïve regressions where censored values were either substituted or omitted from the analysis. Note that the "Truth" fitted line is the true underlying relationship and it is hardly visible because it is overlaid with the censored regression model fit.

(b) The difference between the estimated and true values for the intercept, slope and residual standard deviation used to simulate data for a simulation representing the log-linear relationship between total phosphorus and chlorophyll *a* in north temperate lakes. There were five scenarios evaluated, including a censored regression model and three naïve regressions where censored values were either substituted or omitted from the analysis. Simulations were performed assuming 5% (A), 15% (B), or 30% (C) of the observations being left-censored. The open squares, triangles, and circles represent the mean difference across 100 iterations for the residual standard deviation, slope, and intercept, respectively, and the horizontal bars represent the 2.5 and 97.5 percentiles across the 100 simulations.

Figure 1



Pre 2000
2000s
2010s
All Years
1.00
0.75
0.50
0.25
0.00
0
1
2
3
CHLa (µg/L)
0
100
200
300
TN (µg/L)
0
10
20
30
TP (µg/L)

Figure 2a

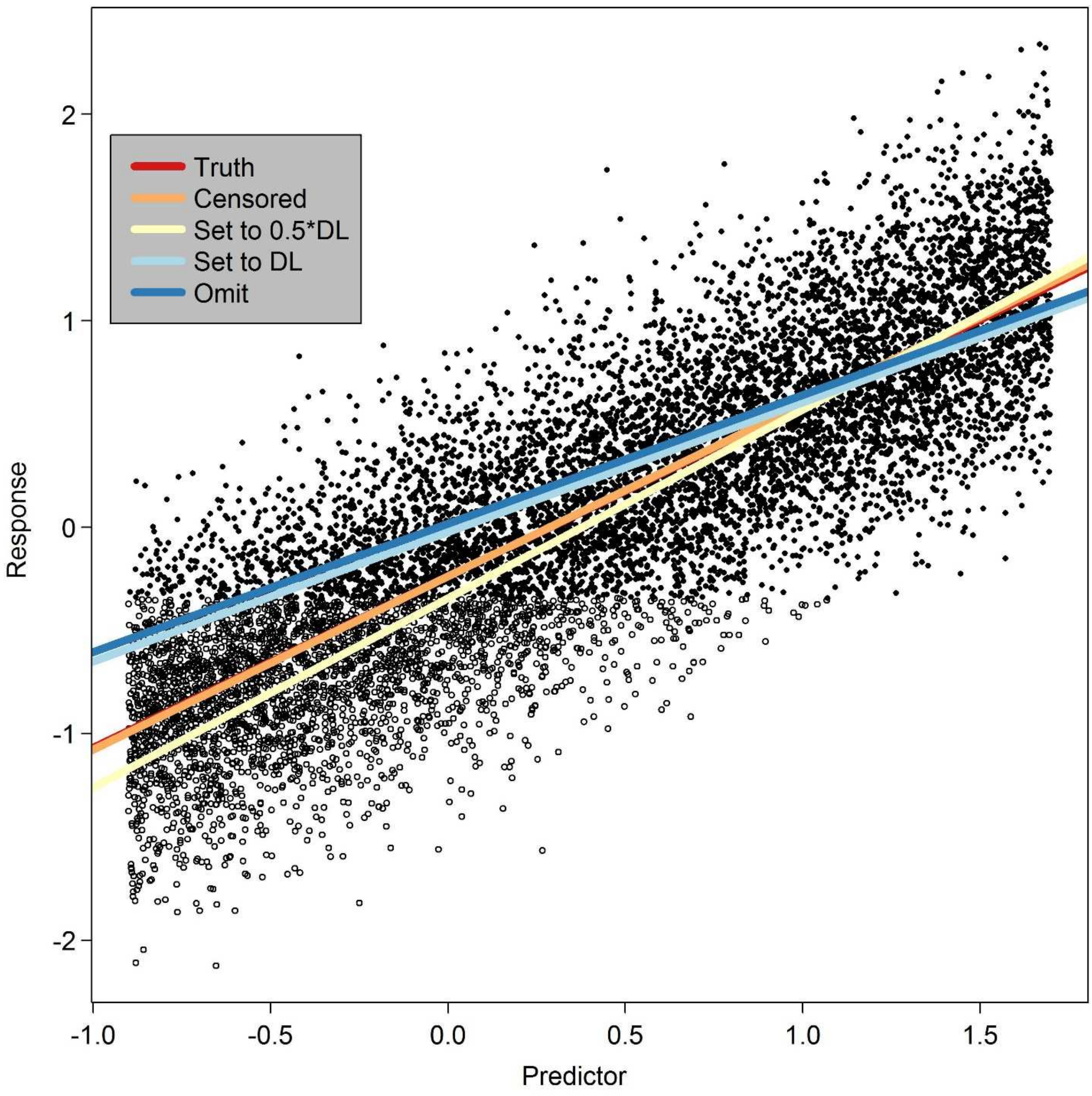
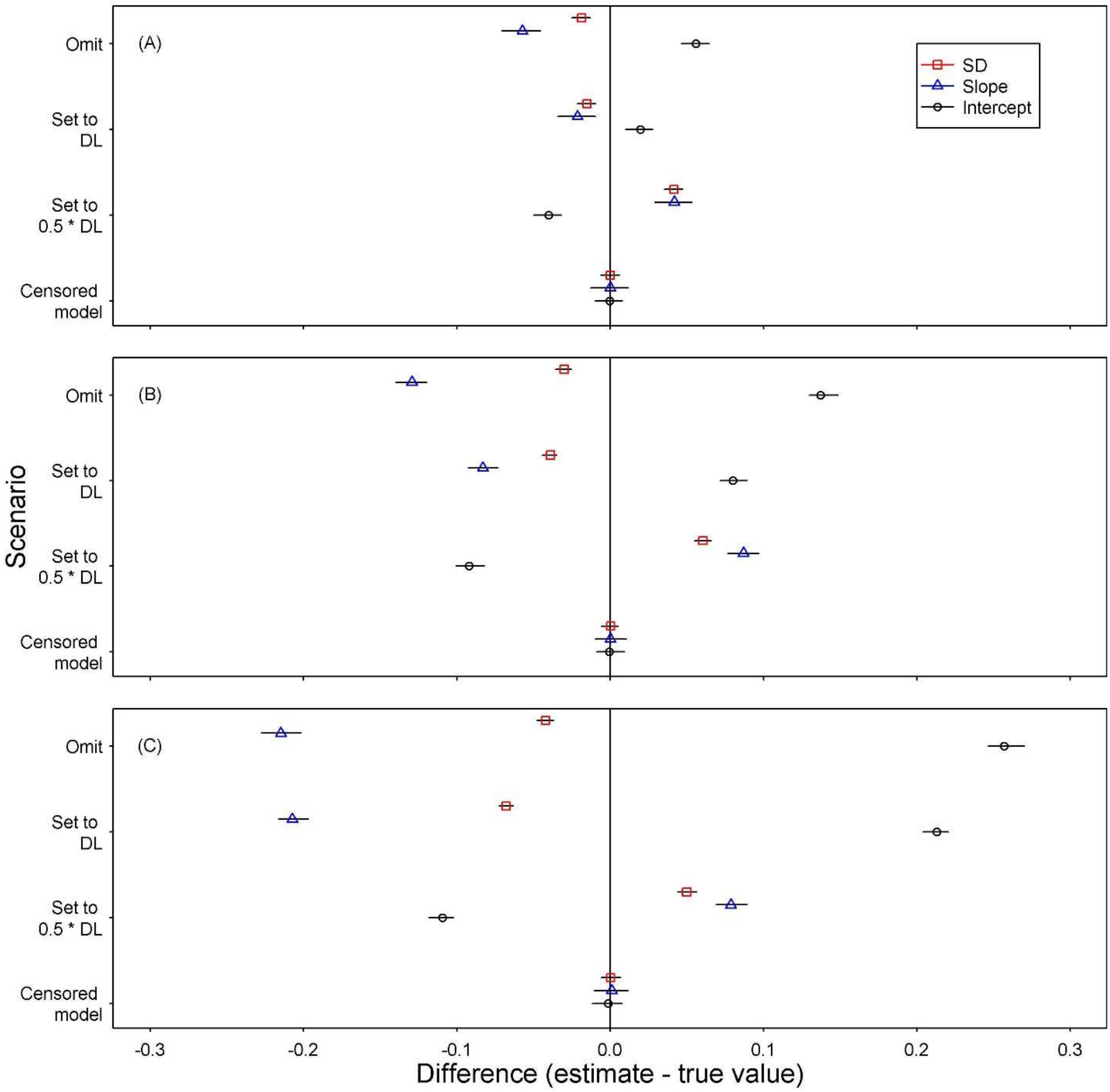

Truth
Censored
Set to 0.5*DL
Set to DL
Omit
Response
Predictor
2
1
0
-1
-2
-1.0
-0.5
0.0
0.5
1.0
1.5

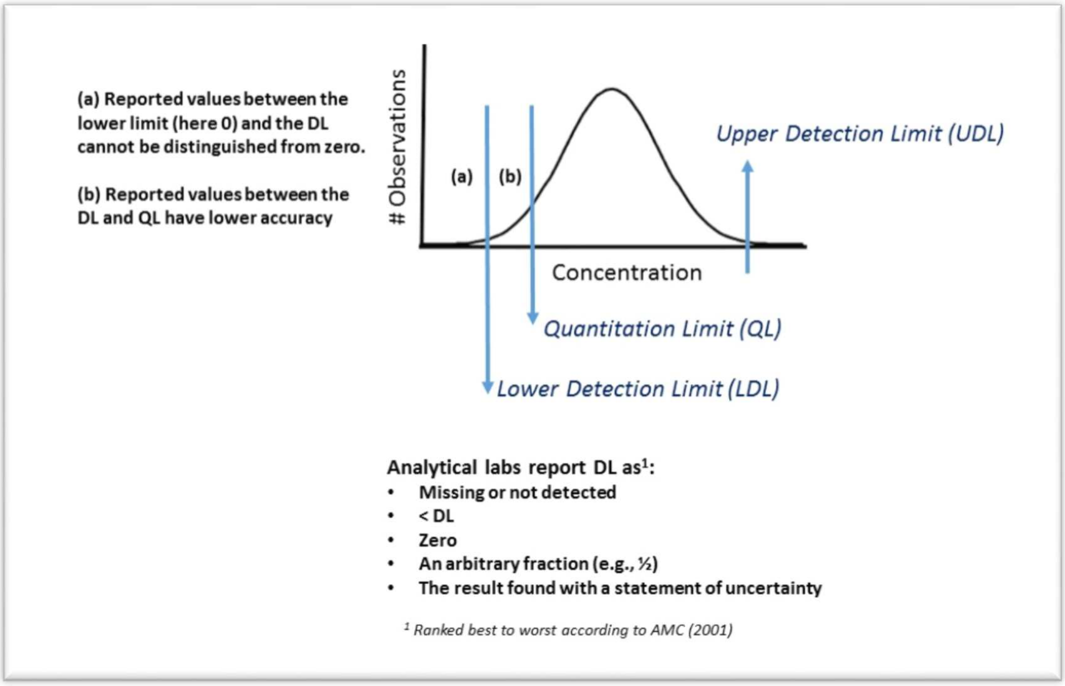
Figure 2b


(A)
(B)
(C)
Omit
Set to DL
Set to 0.5 * DL
Censored model
SD
Slope
Intercept
Scenario
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Difference (estimate - true value)

Supplementary web panel:

Terminology used to define aspects of data quality. Definitions from Helsel (2011).

TERM	DEFINITION
CENSORED DATA	Typically a low level concentration with a value between zero and the reporting limit; can also be a concentration above an upper threshold set by analytical constraints
REPORTING LIMIT	Concentration above which values are reported without qualification by either detection or quantitation limits
DETECTION LIMIT (DL)	Value below which a concentration cannot be distinguished from zero. Related terms are LOD (limit of detection) and MDL (method detection limit)
QUANTITATION LIMIT (QL)	Value below which a reliable single number cannot be reported with precision. Related term is LOQ (limit of quantitation)
DATA SUBSTITUTION	Replacement of censored data in a dataset with, for example, zero, ½ the detection limit, or the detection limit.
TAG OR QUALIFIER	Field in a database that indicates whether a value is censored
MISSINGNESS	The manner in which data are missing from a sample of a population, which can cause artifacts in data analysis under certain conditions



Analytical Methods Committee (AMC) (2001). What should be done with results below the detection limit? Mentioning the unmentionable. AMC Technical Brief (No 5): 2.