Published in final edited form as:
J Expo Sci Environ Epidemiol. 2023 July ; 33(4): 610–619. doi:10.1038/s41370-022-00501-1.

Characterizing surface water concentrations of hundreds of organic chemicals in United States for environmental risk prioritization

Risa R. Sayre[1,2], **R. Woodrow Setzer**[3], **Marc L. Serre**[4], **John F. Wambaugh**[3,4]

[1]Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC, 27709, USA.

[2]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, 135 Dauer Dr, Chapel Hill, NC, 27599, USA.

[3]Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC, 27709, USA.

[4]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, 135 Dauer Dr, Chapel Hill, NC, 27599, USA.

Abstract

Background: Thousands of chemicals are observed in freshwater, typically at trace levels. Measurements are collected for different purposes, so sample characteristics vary. Due to inconsistent data availability for exposure and hazard, it is complex to prioritize which chemicals may pose risks.

Objective: We evaluated the influence of data curation and statistical practices aggregating surface water measurements of organic chemicals into exposure distributions intended for prioritizing based on nation-scale potential risk.

Methods: The Water Quality Portal includes millions of observations describing over 1700 chemicals in 93% of hydrologic subbasins across the United States. After filtering to maintain quality and applicability while including all possible samples, we compared concentrations across sample types. We evaluated statistical methods to estimate per-chemical distributions for chosen samples. Overlaps between resulting exposure ranges and distributions representing no-effect

Corresponding author: Correspondence to Risa R. Sayre, sayre.risa@epa.gov.

Contributions

The original concept for this analysis is based on work by RWS. All authors here were working together on another project; the idea for making this a stand-alone manuscript was identified by JFW and RRS. RRS wrote the text of the article, along with JFW. RRS wrote the extraction script. A first idea for the analysis script was written by RWS (although the current work uses different methods). RRS conceptualized and wrote the current analysis script, with assistance from JFW and RWS. MLS and RWS provided guidance on statistical methods. All authors reviewed and approved both the work as it progressed and the final form of the manuscript.

Disclaimer

The views expressed in this publication are those of the authors and do not necessarily represent the views or policies of the U.S. EPA. Reference to commercial products or services does not constitute endorsement.

Ethics declarations

The authors declare no competing interests.

concentrations for multiple freshwater species were used to rank estimated chemical risks for further assessment.

Results: When we apply explicit data quality and statistical assumptions, we find that there are 186 organic chemicals for which we can make screening-level estimates of surface water chemical concentration. Of the original 1700 observed chemicals, this number decreased primarily due to a predominance of censored values (that is, observations indicating concentrations too low to be measured). We further identify 423 chemicals where all measurements were censored but, through consideration of detection limits, risk might still be prioritized based on the detection limits themselves. In the final set of 1.5 million samples, the median environmental concentration of one chemical (acetic acid) exceeded the 5th percentile of no-effect concentrations for the most delicate freshwater species (the highest priority risk condition identified here), and a further 29 chemicals were identified for possible further evaluation based on a small margin between occurrence and toxicity values.

Significance: This method shows the broad range of chemical concentrations seen for organic chemicals across the country and identifies methods of determining their central tendency, allowing for researchers to characterize higher-than-normal or lower-than-normal surface water conditions as well as providing an overall indication of the presence of organic chemicals in the United States. The highest chemical concentrations did not always indicate the highest-risk conditions. Even when accounting for the high level of uncertainty in these data due to differences in data collection and reporting across the set, some chemicals may still be categorized as higher environmental risk than others using this method, providing value to chemical safety decision makers and researchers by suggesting avenues for more focused investigation.

Keywords

Chemicals; Environmental statistics; Water

Introduction

The potential risk posed by a chemical to public health or the environment may be conceptualized [1] by comparing the threshold at which the chemical has been determined to cause harm in a particular organism to the magnitude (in space, time, or both) of that organism's contact with the chemical (that is, exposure [2]). Yet for thousands of chemicals, neither the harmful dose level nor the expected exposure is known for humans or other species [3, 4]. It would be impractical to develop guideline-quality *in vivo* toxicity data for all species-compound combinations (much less simply for humans), or to measure the concentrations of thousands of chemicals in surface water anywhere contact could occur. Instead, many governmental regulatory bodies are investigating new approaches to identify chemicals most in need of attention among those currently in use [5].

New approach methods (NAMs) for assessing hazard, including using *in vitro* models to measure cell-based toxicity, have become a well-developed field in recent decades and are gaining acceptance for use in regulatory decisions [5,6,7]. NAMs for exposure also exist [8], including high-throughput exposure models [9,10,11], but since modeling exposure is complex (for example, physical transport, chemical transformations, and biological

interactions across different conditions), different modeling assumptions may lead to very different results depending on the underlying framework [12,13,14]. Regardless, decisions on which chemicals in surface water require the most monitoring and regulatory attention are routinely made, implicitly or explicitly, with less-than-ideal information.

The process of screening of thousands of chemicals to identify priorities for follow-up research poses some challenges that differ from the regulation of a specific chemical, even though many of the tools, data, and scientific issues can overlap [13,15,16]. When assessing risk posed by a single chemical, incomplete knowledge may be addressed with a mix of new measurements and simulation of different possibilities based on comparison of similar chemicals [17,18]. Similar resources suitable to address thousands of chemicals are unlikely to be available, so decision makers must make the best-available use of existing data sources and models [19].

For a variety of reasons (such as required routing monitoring or monitoring for a specific research project or in response to water quality questions), the U.S. Geological Survey, the U.S. Environmental Protection Agency (U.S. EPA), state agencies, and other stakeholders routinely collect a range of data, including surface-water and groundwater samples [20]. These data, presented in aggregate form through the U.S. National Water Quality Monitoring Council's Water Quality Portal (WQP) [21], do not necessarily compose a representative survey of the waters of the United States, but they do provide millions of observations with the potential to inform chemical risk prioritization. An additional challenge is that metadata needed to contextualize these samples are often missing or incomplete, which forces analysts to make assumptions [22].

Here we developed chemical-specific estimates for ranges of surface water concentrations that may be compared to estimated hazardous concentrations in order to prioritize environmental chemicals based on the risk posed to human or ecological health on the scale of a nation, in this case the United States. We investigated the effect of different data curation and analysis assumptions on the inference of national-scale chemical concentration distributions. Public health researchers looking for spatiotemporal associations between environment and health outcomes often run into data gaps impeding analysis; if these heterogenous surface water observations can be integrated on a per chemical basis into national-representative values, we might broaden the number of chemicals that can be investigated for health effects.

Materials and Methods

Data were analyzed in Python 3.6 using packages numpy, pandas, os, xml, and pickle; and in R 4.0.5 using libraries EnvStats, tidyr, dplyr, ggplot2, readxl, NADA2, forcats, viridis, cowplot, stringr. All data and analysis scripts have been made available at: https://github.com/USEPA/EcoSEEM-Consensus-Model-for-Chemicals-in-Surface-Water.

Data set

The U.S. National Water Quality Monitoring Council's WQP [21] (https://www.waterqualitydata.us/) includes measurements of surface water and groundwater

samples [20] collected by the U.S. Geological Survey, the U.S. EPA, state agencies, and other stakeholders. Read et al. [21] describe the WQP as "…the largest standardized water quality dataset available at the time of this writing, with more than 290 million records from more than 2.7 million sites in groundwater, inland, and coastal waters."

Data curation

All records for water samples of organic chemicals in the contiguous United States in the WQP from 2008 to 2018 were downloaded and transformed using a Python script available on GitHub (folder observation_data). Table 1 provides definitions for key terms for describing the water samples.

Chemicals were often identified only by name in this database, which may not be sufficient for positive identification [23]. Names and Chemical Abstracts Service Registry Numbers (CAS-RN) were provisionally mapped to unique chemical substance identifiers (DSSTox Substances IDs) using the batch search function in the U.S. EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard [24]); the mappings were then manually confirmed. Chemicals without a match were checked for typos and alternate spellings (load_water_data.py at GitHub). Some chemical names were aggregated to a single identity based on identical structure. Because ecotoxicity values were only available for substances with unique structures, records for chemical names representing ambiguous structures were removed from our dataset.

Once chemical identities were harmonized, records were pruned based on spatial location (for example, samples suspected to be taken from saltwater) and metadata (for example, results not from ambient water monitoring, such as "initial dilution zone", "finished water" or "field spike"). Concentration units were standardized to μg/l for measured concentrations and for limit values (load_water_data.py). For this evaluation, we excluded chemicals with fewer than 50 observations above the detection limit across the 10-year range to avoid inference on the most data sparse chemicals.

Intra-chemical comparison of representative values for different sample types

Recognizing that samples were collected under many circumstances for many purposes (routine monitoring, monitoring for a specific research project or to track reports of contamination), we attempted to identify factors that could cause concentrations to differ. The first data subset we recognized is limit types. Over 80% of the samples in this set were censored: identified only as being below a limit value. We divided the limits into three conceptual categories: (1) a "technical min" representing the lowest possible result detectable using a particular instrument or method (also called a non-detect), (2) a "technical quant" representing the lowest result deemed quantifiable using a particular instrument or method, and (3) a "reporting min" representing the lowest value reportable based on a laboratory's certification or other regulatory threshold. A "technical quant" has a numerical relationship to "technical min", but a "reporting min" is often unrelated to either technical measure. Additionally, for some chemicals and methods, the "technical quant" can vary based on the sample. It can be imagined that samples expected to be high could be measured using a less sensitive method or only reported if over a regulatory limit. Even if our

investigation could not directly answer whether below-limit values were missing at random or not, it could tell us if the uncertainty introduced by using a higher limit could affect a distribution, and therefore a risk prioritization, using the resulting value. For the comparison of distributions by limit type, we used observed values combined with either the records where the limit was a “technical min”, or the records where the limit was a “reporting min”. Samples were identified as being within one of the three types using a text classification script available in our GitHub (load_water_data.py).

The second data subset was a comparison of bulk and dissolved samples (here called “phase”). Samples may be analyzed in different phases based on the objective of the sampling activity or on the properties of the chemical. For example, chemicals known to adhere to organic matter may be less likely to be detectable in dissolved fractions. Labeling samples as dissolved (having passed through at least a 0.45-micron filter) or bulk (less finely filtered) was determined by text analysis of the method description (details in Python code). We also checked for a log-linear relationship between differences in concentration based on phase with several relevant physicochemical properties as explanatory variables.

The final data subset was a comparison of per-chemical concentrations by hydrological quarter: January–March (a season of soil moisture recharge), April–June (a season of runoff), July–September (a season of evapotranspiration), and October–December (another season of soil moisture recharge), based on knowledge that some chemicals such as pesticides are used seasonally and might therefore have different concentrations per season. However, as different pesticides are used on different crops, and growing seasons vary by crop and location, it may be that these known usage differences are not great enough to elicit a statistical change in a chemical’s concentration distribution between seasons, considered with other factors of variability such as persistence.

In all three cases, difference between subsets was determined using a Peto-Peto two-sample test (a log-rank method appropriate for skewed data [25]) to compare differences between their empirical cumulative distribution functions instantiated in the R package EnvStats [26]. The significance level was adjusted by the sample sizes of the comparison groups, which was sometimes quite different. Different degrees of censoring in the sets could possibly bias the comparisons; however, this was something we could neither prevent nor correct for in this highly censored dataset [27]. For all three comparisons, only some chemicals could be compared; for example, samples of some chemicals were only ever analyzed as dissolved. To determine whether subsets would be considered different for our risk prioritization, the number of chemicals with evidence of difference in their distributions by subset was divided by the number of chemicals without. When the ratios were very small, the subsets could be combined without affecting the outcome of a risk decision; other ratios were considered on a case-by-case basis, incorporating other relevant knowledge.

The authors note there may be other factors that could cause concentrations to differ. The analysis methods described here could be applied to any other factors of interest in other experiments.

Calculation of representative values

Because many sample results were below reporting or detection limits (known also as "left-censored"), we needed to employ a method of developing representative values that meaningfully incorporated those values. The commonly used method of substituting left-censored results with the limit value or half the limit value artificially reduces the variation in the sample and introduces bias in the mean [28, 29]. Therefore, we evaluated the difference in representative values calculated using either a parametric method that assumed a log-normal distribution of the data (maximum likelihood estimation (MLE)) or a non-parametric method based on the empirical cumulative distribution function (Kaplan-Meier (KM)). Values for both methods were calculated using the R package EnvStats.

The benefit of MLE is that it allows determination of parameters describing a distribution based on maximization of the likelihood of the detects, while also incorporating knowledge of the non-detects. This permits measures of central tendency and dispersion to be calculated with the below-limit values meaningfully included. However, it assumes that the underlying data were lognormally distributed, while environmental concentrations may in truth be even more right-skewed (concentrations near or at zero, with a few high concentrations) [30]. KM avoids this potential misrepresentation by being strictly empirical; however, because it does not impute values below the smallest observed value, a central tendency derived from a highly censored set will be biased [31]. Although the representation is empirical, in the case of high censoring, there may not be an observed middle value; the position of the KM median was calculated using the Michael and Schucany method as recommended in EnvStats documentation. Therefore, we were able to derive different representative values using these two methods (KM and MLE) to take advantage of their relative strengths and allow analyses of risk from either expected concentrations (based on the medians) or from the highest concentrations (based on the 95th or 99th percentiles).

Inter-chemical comparison of bioactivity and exposure

To characterize bioactivity for comparison to these exposure values, we first queried EPA's ECOTOX Knowledgebase [32] for in vivo-measured toxicity measurements in freshwater species. However, consistent toxicity values were not available across our chemical space. Therefore, we adopted species sensitivity distributions (SSDs) developed by Posthuma et al., which estimated a range of predicted no-toxic-effect values across species for over 12,000 chemicals [33]. Compounds were mapped to DTXSIDs using the CompTox Chemicals Dashboard by name and CAS-RN; mismatches were manually resolved. We regarded chronic concentration distributions as most relevant for this assessment, and only used values where authors deemed there was enough information to completely characterize a distribution (as indicated in their results by quality codes beginning with a "1"). We began risk prioritization by comparing the smallest no-effect concentrations in each distribution (representing concentrations at which no effect is predicted even for a chronic exposure in the most sensitive species) in surface water and the highest occurring concentrations (empirical or estimated).

For chemicals for which no representative value could be determined due to a lack of above-limit values, the highest per-chemical limit value was compared to the bioactivity value

range. Although this comparison holds less information than the exposure distributions, it may still be the case that some per-chemical limit ranges are lower than all hazard values, constituting a no-risk condition based on this data. Prioritization was performed with categories similar to the set for which distributions could be derived, though in this case since only the highest limit was used as an exposure proxy, the categories (and results) between the two types of prioritization are not equivalent because the highest actually-occurring concentration for the second type of comparison may be many times smaller than the highest limit, whereas in the first comparison type the highest value was always a detect; limit values above detects were not incorporated.

Results

Qualitative data curation

Figure 1 shows an overview of how the dataset was winnowed to yield the final analysis set of 1.5 million samples. In our workflow all chemicals were processed consistently, and the amount of information per chemical dictated what could be calculated from each chemical's data. The first step, associating the samples with specific chemicals identities (labeled "Chemicals" in Fig. 1), was executed manually. Comments on chemical identity decisions are provided in the supplement on GitHub load_water_data.py. The second ("Sites") and third ("Samples") sections were filtered by machine, the details of which are available in the same script. It is worth noting that records missing metadata in site or sample fields were not excluded from this analysis. Although those records may include irrelevant data, there was no clear evidence either way, so we retained the data. Therefore, some non-ambient or non-surface water samples may have inadvertently been included. It may be possible to infer this information by text analysis of other metadata fields in a more detailed filtering approach, if necessary.

The number of chemicals (or chemicals' names) is indicated in parathesis, to demonstrate how the number of chemicals available for prioritization decreased because of sample inapplicability or missing information. In the Chemicals stage, "Names represents multiple structures" compounds were aggregated by DTXSID into a category on the left; therefore, a name may be in more than one category. "Result" means a sample had a non-censored value in the record, "LOQ" means a sample had a limit value in the record, and "joined LOQ" means a limit value was found in an analytical method linked to the record.

Influence of water sample characteristics

For each chemical our goal is to determine a nation-wide distribution from which the available samples might have been drawn. Because the samples were collected inhomogeneously and annotated inconsistently, we examined the impact of different approaches to aggregating the data across key sample characteristics:

Annotation of reporting/technical limits

Since many samples are non-detects, it is critically important to understand the lowest detectable limit for each sample. Per chemical, there may be several values for each type of limit (for example, differing instruments yield different technical limits for the same

chemical). Figure 2 shows that for 59% of evaluation set chemicals, the median reporting minimum was higher than the median technical minimum. Almost all chemicals had cases where a limit was reported instead of a measurement. We therefore compared inter-chemical distributions of samples where metadata provided either “technical mins” (51.5% of all samples) or reporting limits (33.9% of all samples). Of the 177 chemicals in the overall set with both “technical mins” and reporting limits, the intra-chemical distributions were statistically different for 131 (using the same above-limit values). We did not expect the difference in limit types to be explicable by evaluating the available metadata, and we do not have a proposed reason for the difference between chemicals with similar distributions across limit types and chemicals where the distributions change. Although it is possible that high concentrations above most observations (but below the even-higher limits) are present, there is no evidence of them in the quantified values. We chose not to exclude samples with reporting limits, as a comparison could only be made between reporting and technical limits for 36% of chemicals. It is possible that the reporting minima add more noise than signal (for example, reporting limits are sometimes several orders of magnitude higher than any observed value) but for this analysis, these records were retained to maintain a greater sample size and a slightly larger chemical space. It could be possible for distributions including non-detects with only reporting limits to skew higher, but the statistical methods used here do not incorporate limits above any observed value when calculating representative values.

Scatter plot showing the per-chemical median of quantified samples for the evaluation set chemicals on the *x*-axis and the per-chemical median of limits on the *y*-axis, with reporting limit medians as blue squares and technical limit medians as red triangles. The dashed line represents a 1:1 relationship, not a regression line.

Bulk vs. dissolved water samples

We next examined the impact of the phase (bulk or dissolved) of the samples on intra-chemical distributions. The phase of each sample was identified using three different metadata fields in the ResultAnalyticalMethod table from the WQP (details in Python script). Across a test of 800,000 records, no phase conflicts were identified by our script using any of those fields, therefore we assumed an entry in any field was sufficient to determine the phase of a sample. However, for about 21% of samples, either all three fields were blank, or the field text was not categorized into a phase by our identification script; these samples were not included in the final calculations. When comparing the intra-chemical distributions for the 272 chemicals that had both bulk and dissolved samples, 165 of them had statistically different distributions at a significance level of 0.05. Although phase distributions were more similar than limit distributions, aquatic toxicity values are measured in dissolved concentrations. If the bulk and dissolved concentrations were similar, they could be kept together to increase the sample size, but since notable differences were observed, only the samples defined as dissolved (46%, or 2,292,937 samples) were used for this risk prioritization. This led to the removal of 24 chemicals that were only measured in bulk or unknown phases. When upholding our constraint of chemicals with at least 50 observed values, the set was reduced to 286 chemicals for the rest of the analysis. In a logistic regression based on a binomial distribution on the physicochemical properties

air:water partition coefficient, octanol:water partition coefficient, vapor pressure, and water solubility (predicted using the OPERA model [34]), there were no significant predictors of whether bulk and dissolved concentration distributions would differ.

Seasonal variation

For the final intra-chemical comparison, concentration distribution differences by season, there are six possible pairs (for example, one comparison is Jan–Mar compared with Apr–Jun). The distribution across chemicals for the number of statistically different season-season pairs is shown in Fig. 3. Although there are three or more season pairs with concentration differences for most chemicals, which season-pairs differed varied across chemicals. The most hydrologically similar season pairs (Jan–Mar and Oct–Dec) had the fewest differences across all chemicals to a small degree (47%), but none of the season pair distributions were as different from each other as were the bulk and dissolved concentrations. To test the hypothesis that pesticide active ingredients would have more dramatic seasonal differences than other chemicals, we compared the count of season pair differences between pesticides (as defined in the EPA Chemicals Dashboard list EPAPCS: https://comptox.epa.gov/dashboard/chemical-lists/EPAPCS) and non-pesticides. As shown in Fig. 3, the distribution of differences is similar when comparing pesticides to other chemicals. This may imply that hydrological differences (for example, seasonal changes in rainfall) are more influential on ambient national concentrations than usage patterns, but making that determination was out of the scope of this project. None of the seasons had the highest or lowest concentrations across chemicals, and sampling activity, while present in all seasons for many chemicals, was very uneven across seasons. There were more than twice as many samples taken in the second and third seasons as in the first and fourth seasons, which influenced the certainty of the values and added bias to comparisons of sample sets. Therefore, we decided to include samples from any season in our estimates as we felt differences would be better addressed by a model able to incorporate data on hydrological differences between seasons.

This figure shows the histogram of differences between per-chemical concentrations on the basis of per-season averages; with zero representing statistically similar concentrations in all season comparisons. Given four seasons, a count of six represents statistically different concentrations for all possible season-season comparisons for a given chemical. The pesticides (shown below) have a similar distribution of differences to the non-pesticides (shown above). There are a few more pesticides than non-pesticides with statistically different concentrations in all season-season pairs, but this may be a data artifact because there are more pesticides overall in the set than non-pesticides. Clear seasonal concentration patterns across chemicals were not observed in this set.

Calculation of representative values

Once the method for curating acceptable data was selected, the most salient challenge was how to address the large proportion of non-detects (illustrated in Fig. 4). Referring to Fig. 2, just as the reporting limits are usually higher than the technical limits, the technical limits are still higher than the median result values for 66% of chemicals. And as shown in Fig. 2, limit values (the first sample type comparison) were correlated with the measured

value. There was a near 1:1 ratio for the median observed value and the median limit value in this dataset, a source of measurement uncertainty for any study of ambient water concentrations. Analytical chemists may use methods with sensitivity commensurate to the values being measured but as previously discussed, meaningful incorporation of limit values also is confounded because not all limits relate to a technical measure: perhaps reporting limits also relate to expected concentrations in some cases. Limit values that exceeded the highest observed value for a chemical were not incorporated, and led to another source of data loss. Our estimated values are summarized in all_chem_res.csv (GitHub).

Scatter plot showing the record count for each chemical in the final analysis set on the *x*-axis and the percent of those records with only left-censoring limits instead of measurements on the *y*-axis.

Figure 5 shows representative values calculated using two different methods for developing a distribution given censored data, KM and MLE, for all chemicals within the final evaluation set. For the MLE method was the estimated mean of the log-normal distribution —which is the median of the concentrations—was used as the representative value, while for KM the estimated median was used. In general, the two methods had the greatest concordance for chemicals with lower censoring proportions. Referring to the long-dashed bars on the right of the graph, notice that the variance is large in MLE for highly censored chemicals, pushing those medians lower than those calculated using KM. Using median-only estimates in these cases could possibly lead to falsely low values, but it could also be true that in some samples, the chemical analyzed was not present at all. It was not possible to tell from available data which scenario was the case. For 22 chemicals parameters for a distribution could not be estimated through MLE with the EnvStats function eqlnormCensored due to a wide spread of observed values or other data irregularities. We noted that the KM estimates were sometimes lower than the MLE estimates in cases of high variance.

Comparison of concentration ranges (median to 95th percentile) using KM (dotted) and MLE (dashed), with chemicals arranged from least censored on the left to most censored on the right. The highest censoring level is shown with an upward pointing triangle; these values are usually higher than all observations.

When examining the concentrations, four of the ten chemicals with the highest median concentrations were listed as pesticide active ingredients. However, some of these pesticide active ingredients such as di(2-ethylhexyl) phthalate also may have entered the environment from industrial or post-consumer releases unrelated to pesticides. Within these top ten were also disinfection byproducts (bromodichloroacetic acid and dichloroacetic acid) and breakdown products (acetic acid and methane) that could come from many types of sources, naturally occurring and human made. When inspecting the highest 99th percentile concentrations, six of the ten chemicals are listed as pesticide active ingredients; eight chemicals are in both lists. Depending on the risk assessment or research interest, the chemicals that would motivate further research would differ. The presence of high concentrations of disinfection byproducts in surface water, for example, could have human health implications if they are not removed efficiently during drinking water treatment

processes. The results of this research can identify avenues for research and the methods can inform the way relevant data for the research is recognized and analyzed.

People and wildlife are commonly exposed to chemicals through water, so the representative values calculated here might provide a point of reference for interpreting the national context of a given sample. There are many reasons the concentration of a chemical might be higher than the representative value reported here; for example, our estimates are obtained by averaging spatially over the entire United States and therefore proximity to a chemical release might reasonably lead to exceeding our estimated value. For example, for Di(2-ethylhexyl) phthalate, the surface water concentration would have to be above 0.2 ug/l to be above the median level we estimated, possibly indicating a source nearby (among many other reasons). Meanwhile for Pyraclostrobin, the surface concentration would only have to be above 0.00137 ug/l to be above the average surface water value estimated here. For 661 other chemicals, such as Hydrocortisone, measurements were attempted in more than 50 surface water samples, but the concentration was above the limit of quantification in fewer than 50 of all samples. However, the median censoring level (that is, detection limit) for Hydrocortisone was 0.147 ug/l, and therefore any concentration measured above 0.147 ug/l might indicate a proximate source.

Bioactivity:exposure ratio comparison

We next compared the environmental occurrence ranges with hazard as represented by the Posthuma cross-species distributions of no-toxic-effect concentrations (that is, SSDs) [33] as an example national screening-level risk comparison. After removing the SSDs with the lowest confidence flags (indicating insufficient evidence for developing a distribution), there were 186 chemicals remaining for a bioactivity exposure risk prioritization (Fig. 6). Because the KM values were more closely tied to the data and slightly higher (and therefore presumedly more health-protective), the KM 95th percentile values were used for the environmental occurrence part of our bioactivity exposure comparison. This approach also had the benefit of retaining all chemicals for risk prioritization. We developed several categories of risk conditions from higher risk (cases where species encounter predicted no-effect concentrations at the median of the occurrence distribution), to lower risk (no-effect concentrations at the 95th or 99th percentile of the occurrence distribution), to essentially no risk according to this assessment method (no overlap between hazard and exposure distributions). Of the evaluated chemicals, 156 fell into this no-risk category. Of the remaining chemicals, the highest-risk condition with any representation (in which the exposure median exceeded SSD 5th percentile) contained only one chemical (acetic acid). Although this chemical is relatively non-toxic, it had high levels of observed occurrence. For eight chemicals, the exposure median exceeded the 1st percentile SSD. Most of these chemicals have low environmental concentrations but their smallest predicted no-effect concentrations were even lower. The chemicals listed in Fig. 6 are ordered from left-to-right based on highest risk category to lowest, however since both quantities were uncertain, the absolute ranking of chemicals remains uncertain.

All 186 chemicals with estimated surface water concentrations are shown in the first panel (above), while only those with bioactivity exposure overlaps are shown in the second panel

(below). Each line pair represents one chemical, with the dashed lines displaying the median to 1st percentile of no-effect concentrations in the species sensitivity distribution in log-µg/l and the dotted lines displaying the median to 99th percentile of the KM distribution of dissolved organic chemical concentrations. The highest risk category (left side of figure), the case in which the KM exposure median exceeds the SSD 5th percentile, contains one chemical. The second category, the case in which the KM exposure median exceeds the SSD 1st percentile contains eight chemicals. From left to right, the same two SSD percentiles are used but the overlap with exposure values goes to the 95th, and then the 99th percentile. For 156 chemicals there is effectively no risk based on this evidence; the entire range of hazard values is higher than all observed concentrations). The first part of the figure displays results for all chemicals and the second part of the figure displays the same information zoomed in for the chemicals with any overlap.

For chemicals where there were at least 50 samples but not 50 above-limit dissolved results (423 chemicals, 164 of which had toxicity values) the highest technical or reporting limit value was compared to the hazard distribution—presumedly the mean chemical concentration is less than this limit. For 100 cases even the highest limit value (which was unlikely to be an actually-occurring concentration) was below the 1st percentile SSD. This could be considered as a no-risk condition (based on this dataset), as even the largest upper-limit in the set for that chemical did not reach the level predicted to cause no effect in the most sensitive species. This is even more stringent than the no-risk condition in the other comparison, though it is based on fewer records. In the next comparison category, 14 chemicals had a highest upper-limit above the 1st percentile SSD and below the 5th; in the next category between the 5th percentile and the median of the SSD there were 17 chemicals. In what was considered the highest risk condition in this analysis, 33 chemicals had a higher upper-limit than the median SSD. It is still entirely plausible that even these chemicals considered as being in the highest risk category have their lowest actually-occurring environmental concentrations below all levels expected to elicit a biological effect across species, but due to non-reported concentrations that level of specificity cannot be known from this set.

Discussion

Water is an important source of chemical exposure for both humans and ecological species. We have organized, filtered, and analyzed several million observations reported by the U.S. National Water Quality Monitoring Council’s WQP to estimate median and 95th percentile organic chemical concentrations at a national-scale. Although there were many measurements available in this resource from across the United States, they were collected for different purposes with different methods, and many aspects of the samples vary. The approach we used trades chemical-specific precision for broader (that is, across many chemicals) applicability. By attempting to characterize the uncertainty and limitations of these estimated surface water concentrations, we hope that decision makers may, in some cases, find these summary data fit for screening-level purposes. The workflow could be modified or tailored to evaluate specific geographical areas or sample type subsets.

Of the 1761 chemical names among the observations available from WQP, we could only complete risk-based prioritization (requiring both hazard and exposure estimates) for 186 compounds. Although additional sampling would improve the chemical space of such an analysis there are other simpler things that could also improve the size of the evaluation set. For example, hundreds of the chemical names were ambiguous. Given the level of effort taken in collecting and measuring these samples, a low marginal investment into more detailed annotation, such as providing an unambiguous chemical name, could potentially provide a good return on the ability to use the resulting data for this or other purposes. We hope that by demonstrating that this dataset might have uses beyond the original reason for collection, we provide the motivation for improved annotation and curation of these data. However, we have demonstrated that even when inadequate values are reported to develop estimates of environmental exposure, just reporting the limits can provide enough information to delineate possible risk strata for chemicals.

The largest reduction in chemical space (from 1310 to 498) occurred for chemicals lacking 50 or more uncensored observations. For these chemicals, the highest limit value provides a rough but potentially useful estimate of the upper limit for these chemicals, though it is plausible that their actual concentrations might be far lower (file sayre_water_data_results.csv). Choosing a technical limit value rather than a reporting limit value (when there is that option) gives a better sense of the concentration boundary, given how much higher many of the reporting limits are than the observations. However, it was not always the case that reporting limits were higher than technical limits; it is not known whether this reflects differences in sampling over space and time or errors in the method text identification script. Regardless, unclear chemical naming and high reporting limits were the factors most responsible for the loss of nation-wide representative values; addressing either factor has the potential to considerably increase the number of chemicals that can be compared to bioactivities.

In any dataset including non-detects, assumptions made about how to handle the left-censored observations are critical [35]; especially given such a high prevalence of censored data as in this dataset (Fig. 4). Knowledge of the detection limits was inconsistently annotated across the data. For the chemicals where there were enough detects (our assumption was 50 or more), we found that using the different types of limits (median reporting minima vs. technical minima) led to different estimates for 131 of the 177 chemicals where both types of limits were available. The more censored values there were for a chemical, the stronger the dependence of the median estimate on our assumption of lognormality. The occasional difficulties of the MLE algorithm to converge could indicate that the log-normal assumption was not suited for those samples. One alternative distribution is a mixture between a log-normal and a point mass at zero (that is, the chemical is completely absent in some samples).

We found that the dissolved concentration was often, but not always, statistically different from the bulk concentration, although this may only be the case for the subset of chemicals for which concentrations were measured in multiple phases and not a valid assumption across all chemicals. Six pairwise comparisons of measurements made in different seasons were evaluated, with none being as different from each other as the bulk and dissolved

concentrations. We did not find an overall difference in seasonal dependence between chemicals with pesticide and non-pesticidal uses. For the seasonal comparisons, the number of measurements available in each season was a confounder. For all intra-chemical comparisons, an additional issue was that the statistical tests used assumed independent samples. All the observations were spatially dependent and had other aspects that almost guaranteed that some samples were related to each other (for example, samples taken by different agencies). This analysis did not consider many other factors that could influence overall risk, such as pH or combined exposure to multiple chemicals with a shared mode of action leading to biological changes at smaller per-chemical concentrations. However, it does investigate some of the many possible factors that could be important (but may often be overlooked) when deciding whether to include a sample record in an analysis.

Although the data were selected to be as inclusive as possible, they still only painted a sparse picture of the water concentrations on a spatiotemporal basis. At least one record was present in 2114 of 2270 hydrologic subbasins in the United States in the original set; however, after filtering, there were samples in only 1197 subbasins. The average number of subbasin samples for a given chemical was 280. Temporally, only 340 chemicals were measured in all monitoring years at any site. Future work is needed to evaluate methods for identifying spatiotemporal trends and space-time varying the likelihoods of chemical concentrations in water in the United States.

There were 186 chemicals remaining for the bioactivity exposure risk prioritization. The choice of quantiles used from the exposure and hazard distributions reflect the degree of conservatism of the prioritization. For chemicals with overlapping hazard and exposure, some had high exposure values but relatively low toxicity (acetic acid) and some had very low environmental concentrations predicted to possibly reach a no-effect level in sensitive species under chronic exposure (bifenthrin). In either case, the idea that a national-level concentration might be causing biological effect is an indication of need for further study. Because of the uncertainties involved (as evidenced by the data gaps described here as well as inconsistent sampling availability), it is reasonable to expect that with additional analysis (and potentially new data) that there could be a margin of safety between even these chemicals. It is worth noting that even though some chemicals were identified as a potential ecological risk for the United States, this method does not address chemicals which may exist in environmental concentrations exceeding activity values in different scales of space and time. A similar experiment recreating the tests done here could be done at other scales to determine if the results found in our comparison of different sample types hold true under differing problem scopes.

In addition to chemical exposure and risk prioritization, it is hoped that nation-scale estimates of surface water concentrations based on this set may also serve as evaluation data for predictive chemical fate models for water concentration of other chemicals where no such data exist. High-throughput models help fill data gaps by making predictions largely from chemical structure-derived properties [8]. By developing estimates for a broad range of chemicals, we may better identify when, and how, these high-throughput models can help inform human and ecological risk decisions and, conversely, for which classes of chemicals the models may currently be inadequate. The results of our case study also reiterate the

importance of including both per-chemical exposure values and toxicity values in risk estimates, as the highest concentration chemicals did not always present higher relative risk and lower concentration chemicals sometimes presented higher relative risk in cases where they were potently hazardous for the included range of species.

These data do not result from a randomized sampling scheme but might offer an evocative glimpse of ambient water concentrations in the United States. By contrast, the CDC NHANES chemical exposure human biomonitoring program makes use of carefully chosen individuals for which statistical reflection of the aggregate U.S. population is known. Those data are then analyzed in a standardized fashion over a limited amount of time. The water concentration data analyzed here were collected for myriad reasons using multiple chemical analysis laboratories across decades. The sample locations and timing may reflect external factors such as concern for nearby sources of potential chemical emission. Therefore, the values might be presumed to be conservative and more likely to represent high values, yet we see in Fig. 2 that the distributions are highly skewed toward non-detects. On the other hand, given the paucity of environmental sampling in general (Fig. 4, [4]), we cannot presume to know what has not been observed.

The WQP data used in our analyses do not tell a complete story of national water concentrations, and various problems with the annotation preclude much of the data being included in a national-scale aggregate analysis. However, with reasonable assumptions and thorough data curation, concentration ranges useful for screening-level risk assessment can be estimated for nearly 200 organic chemicals. Additionally, we demonstrate that screening-level risk assessment is possible even when calculation of a concentration range is not possible due to incomplete reporting; a further 164 chemicals may be included in such assessments using the approach described here. The ranges can be used not only for the ecological risk estimate example presented here, but also other purposes such as comparing concentrations at a given location with a national range for geographical epidemiology studies examining a correlation between a chemical and a disease prevalence to determine whether concentrations at different locations are relatively high or low compared with nation-wide values, for example. Another possible use is to investigate chemicals with the highest national concentrations in surface water (which could differ based on whether the median or 95th percentile were used) for potential human contact or risk via recreation or drinking water. So the whole proves to be bigger than the sum of the parts. While we cannot prove that the method for summarizing surface water concentrations described here was the best possible method, we hope by articulating our approach and demonstrating its utility, we draw attention to the impression that data curation assumptions may have on future analyses of this type.

Acknowledgements

The authors thank Dr. Jon Arnot, Ms. Lindsay Eddy, Ms. Colleen Elonen, and Dr. Peter Fantke for their helpful reviews of the manuscript.

Funding

The United States Environmental Protection Agency (EPA) through its Office of Research and Development (ORD) funded the research described here. This project was supported in part by an appointment to the Research

Participation Program at the Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.

Data availability

The initial analysis set was downloaded from https://www.waterqualitydata.us/ using queries described in the file load_water_data.py, hosted at https://github.com/USEPA/EcoSEEM-Consensus-Model-for-Chemicals-in-Surface-Water/tree/master/observation_data. The representative concentration ranges and bioactivity:exposure ratio results are available at the same GitHub repo in the file all_chem_res.csv.

References

1. National Research Council. Risk assessment in the federal government: managing the process Washington (DC): National Academies Press (US); 1983.
2. Zartarian V, Bahadori T, McKone T. Adoption of an official ISEA glossary. J Expo Anal Environ Epidemiol 2005 Jan;15(1):1–5. doi: 10.1038/sj.jea.7500411. [PubMed: 15562291]
3. Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. The toxicity data landscape for environmental chemicals. Environ Health Perspect 2009 May; 117(5):685–695. doi: 10.1289/ehp.0800168. [PubMed: 19479008]
4. Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, Cohen Hubal EA. The exposure data landscape for manufactured chemicals. Sci Total Environ 2012 Jan 1;414:159–66. doi: 10.1016/j.scitotenv.2011.10.046. [PubMed: 22104386]
5. Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, Thomas RS. Accelerating the Pace of Chemical Risk Assessment. Chem Res Toxicol 2018 May 21;31(5):287–290. doi: 10.1021/acs.chemrestox.7b00339. [PubMed: 29600706]
6. Turley AE, Isaacs KK, Wetmore BA, Karmaus AL, Embry MR, Krishan M. Incorporating new approach methodologies in toxicity testing and exposure assessment for tiered risk assessment using the RISK21 approach: Case studies on food contact chemicals. Food Chem Toxicol 2019 Dec;134:110819. doi: 10.1016/j.fct.2019.110819. [PubMed: 31545997]
7. Parish ST, Aschner M, Casey W, Corvaro M, Embry MR, Fitzpatrick S, et al. An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. Regul Toxicol Pharmacol 2020 Apr;112:104592. doi: 10.1016/j.yrtph.2020.104592. [PubMed: 32017962]
8. Wambaugh JF, Bare JC, Carignan CC, Dionisio KL, Dodson RE, Jolliet O, et al. New Approach Methodologies for Exposure Science. Current Opinion in Toxicology 2019;15: 76–92. doi: 10.1016/j.cotox.2019.07.001.
9. Arnot JA, Mackay D, Webster E, Southwood JM. Screening level risk assessment model for chemical fate and effects in the environment. Environ Sci Technol 2006 Apr 1;40(7):2316–23. doi: 10.1021/es0514085. [PubMed: 16646468]
10. Barber MC, Isaacs KK, Tebes-Stevens C. Developing and applying metamodels of high resolution process-based simulations for high throughput exposure assessment of organic chemicals in riverine ecosystems. Sci Total Environ 2017 Dec 15;605–606:471–481. doi: 10.1016/j.scitotenv.2017.06.198.
11. Rosenbaum RK, Huijbregts MAJ, Henderson AD, Margni M, McKone TE, van de Meent D, et al. USEtox human exposure and toxicity factors for comparative assessment of toxic emissions in life cycle analysis: sensitivity to key chemical properties. Int J Life Cycle Assess 16, 710 (2011). 10.1007/s11367-011-0316-4.
12. Schmolke A, Thorbek P, Chapman P, Grimm V. Ecological models and pesticide risk assessment: current modeling practice. Environ Toxicol Chem 2010 Apr;29(4):1006–12. doi: 10.1002/etc.120. [PubMed: 20821532]
13. Arnot JA, Brown TN, Wania F, Breivik K, McLachlan MS. Prioritizing chemicals and data requirements for screening-level exposure and risk assessment. Environ Health Perspect 2012 Nov;120(11):1565–70. doi: 10.1289/ehp.1205355. [PubMed: 23008278]

14. MacLeod M, Scheringer M, McKone TE, Hungerbuhler K. The State of Multimedia Mass-Balance Modeling in Environmental science and decision-making. Environ Sci Technol 2010 Nov 15;44(22):8360–4. doi: 10.1021/es103297w. [PubMed: 20964363]
15. Mitchell J, Arnot JA, Jolliet O, Georgopoulos PG, Isukapalli S, Dasgupta S, Pandian M, Wambaugh J, Egeghy P, Cohen Hubal EA, Vallero DA. Comparison of modeling approaches to prioritize chemicals based on estimates of exposure and exposure potential. Sci Total Environ 2013 Aug 1;458–460:555–67. doi: 10.1016/j.scitotenv.2013.04.051.
16. Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, et al. High-throughput models for exposure-based chemical prioritization in the ExpoCast project. Environ Sci Technol 2013 Aug 6;47(15):8479–88. doi: 10.1021/es400482g. [PubMed: 23758710]
17. Fryer M, Collins CD, Ferrier H, Colvile RN, Nieuwenhuijsen MJ. Human exposure modelling for chemical risk assessment: a review of current approaches and research and policy implications. Environmental Science & Policy 2006; 9(3):261–274. doi: 10.1016/j.envsci.2005.11.011.
18. Hommen U, Baveco JM, Galic N, van den Brink PJ. Potential application of ecological models in the European environmental risk assessment of chemicals. I. Review of protection goals in EU directives and regulations. Integr Environ Assess Manag 2010 Jul;6(3):325–37. doi: 10.1002/ieam.69. [PubMed: 20821697]
19. Ring CL, Arnot JA, Bennett DH, Egeghy PP, Fantke P, Huang L, et al. Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways. Environ Sci Technol 2019 Jan 15;53(2):719–732. doi: 10.1021/acs.est.8b04056. [PubMed: 30516957]
20. Hirsch RM, Fisher GT. Past, Present, and Future of Water Data Delivery from the U.S. Geological Survey. Journal of Contemporary Water Research & Education 2014;153(1):4–15.
21. Read EK, Carr L, DeCicco LA, Dugan H, Hanson PC, Hart JA, et al. Water quality data for national-scale aquatic research: the Water Quality Portal. Water Resour Res 2017;53:1735–45. 10.1002/2016WR019993.
22. Sprague LA, Oelsner GP, Argue DM. Challenges with secondary use of multi-source water-quality data in the United States. Water Research 2017;110:252–261. [PubMed: 28027524]
23. Grulke CM, Williams AJ, Thillanadarajah I, Richard AM. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. Comput Toxicol 2019 Nov 1;12:10.1016/j.comtox.2019.100096. doi: 10.1016/j.comtox.2019.100096.
24. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform 2017 Nov 28;9(1):61. doi: 10.1186/s13321-017-0247-6. [PubMed: 29185060]
25. Lucijanic M, Skelin M, Lucijanic T. Survival analysis, more than meets the eye. Biochem Med (Zagreb) 2017 Feb 15;27(1):14–18. [PubMed: 28392721]
26. Millard SP. EnvStats: An R Package for Environmental Statistics New York: Springer; 2013.
27. Neuhaus G Conditional Rank Tests for the Two-Sample Problem Under Random Censorship. The Annals of Statistics 1993;21(4):1760–79. http://www.jstor.org/stable/2242315.
28. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 2006 Dec;65(11):2434–9. doi: 10.1016/j.chemosphere.2006.04.051. [PubMed: 16737727]
29. Shoari N, Dubé JS. Toward improved analysis of concentration data: Embracing nondetects. Environ Toxicol Chem 2018 Mar;37(3):643–656. doi: 10.1002/etc.4046. [PubMed: 29168890]
30. Helsel DR, Hirsch RM, Ryberg KR, Archfield SA, Gilroy EJ. 2020, Statistical methods in water resources: Techniques and Methods Reston, VA: U.S. Geological Survey; 2020. 458 p. Report 4-A3. doi: 10.3133/tm4a3.
31. Zhong M, Hess KR. Mean Survival Time from Right Censored Data. COBRA Preprint Series 2009 Dec; Working Paper 66. http://biostats.bepress.com/cobra/art66
32. U.S. Environmental Protection Agency. 2020. ECOTOX User Guide: ECOTOXicology Knowledgebase System Version 5.3. Available: http:/www.epa.gov/ecotox/
33. Posthuma L, van Gils J, Zijp MC, van de Meent D, de Zwart D. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12

386 chemicals. Environ Toxicol Chem 2019 Apr;38(4):905–917. doi: 10.1002/etc.4373. [PubMed: 30675920]

34. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 2018 Mar 8;10(1):10. doi: 10.1186/s13321-018-0263-1. [PubMed: 29520515]
35. George BJ, Gains-Germain L, Broms K, Black K, Furman M, Hays MD, Thomas KW, Simmons JE. Censoring Trace-Level Environmental Data: Statistical Analysis Considerations to Limit Bias. Environmental Science & Technology 2021 Feb 24;55(6):3786–95. [PubMed: 33625843]

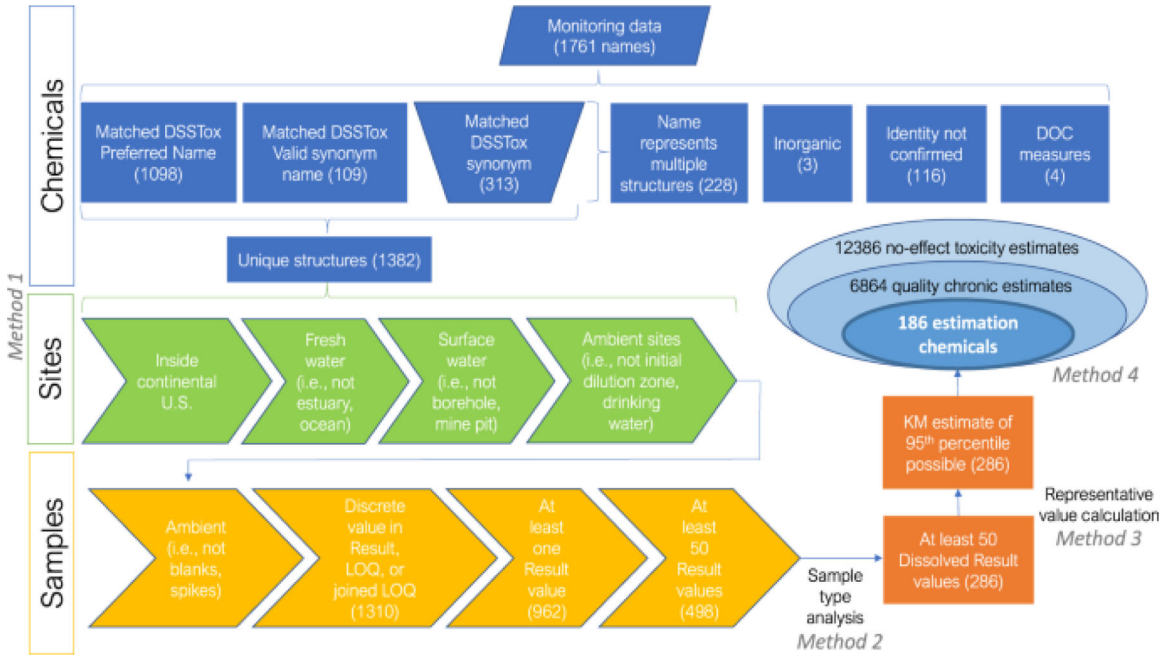

Fig. 1:

Visual representation of criteria used to filter samples and arrive at the final evaluation set.

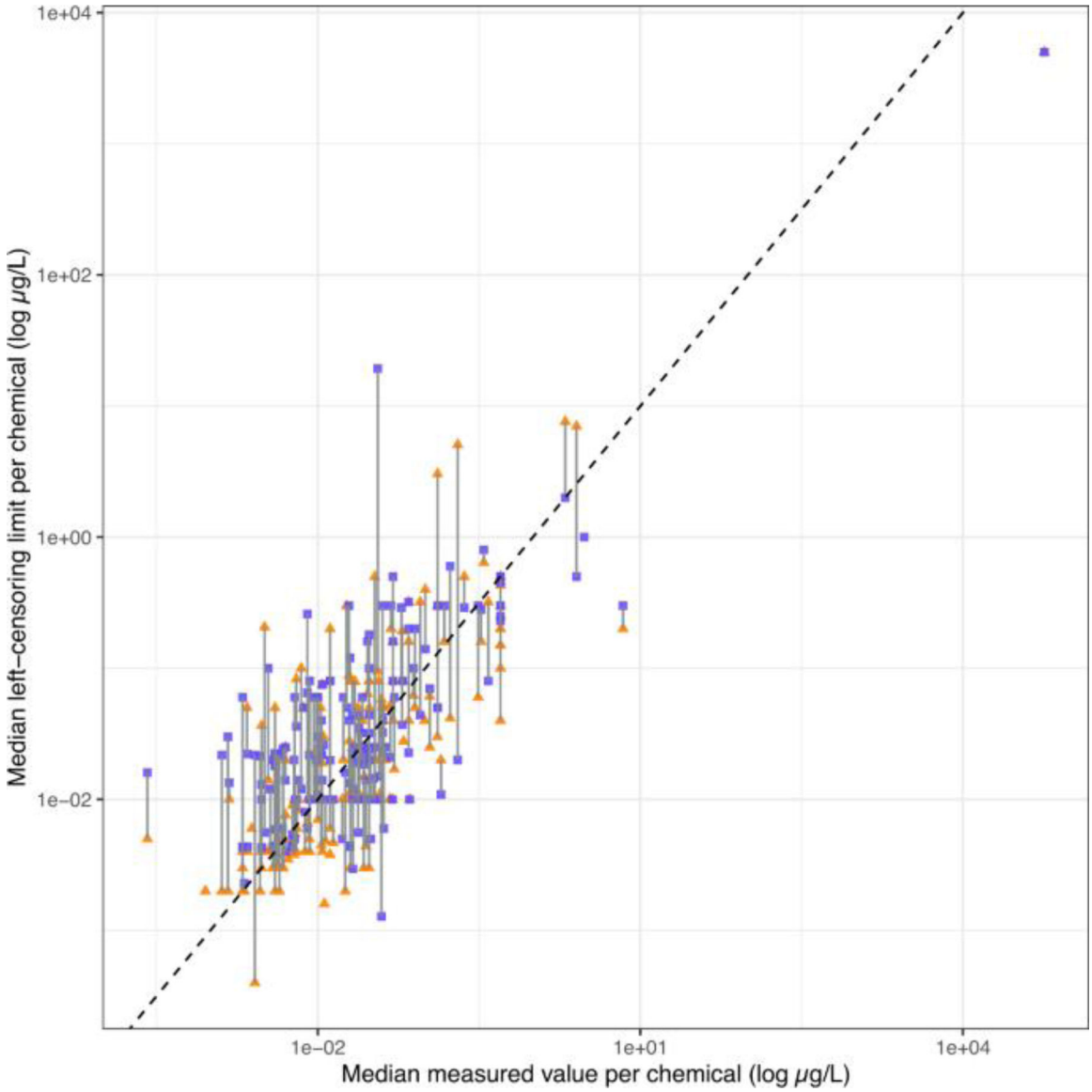

Fig. 2:
Limit values often exceed result values.

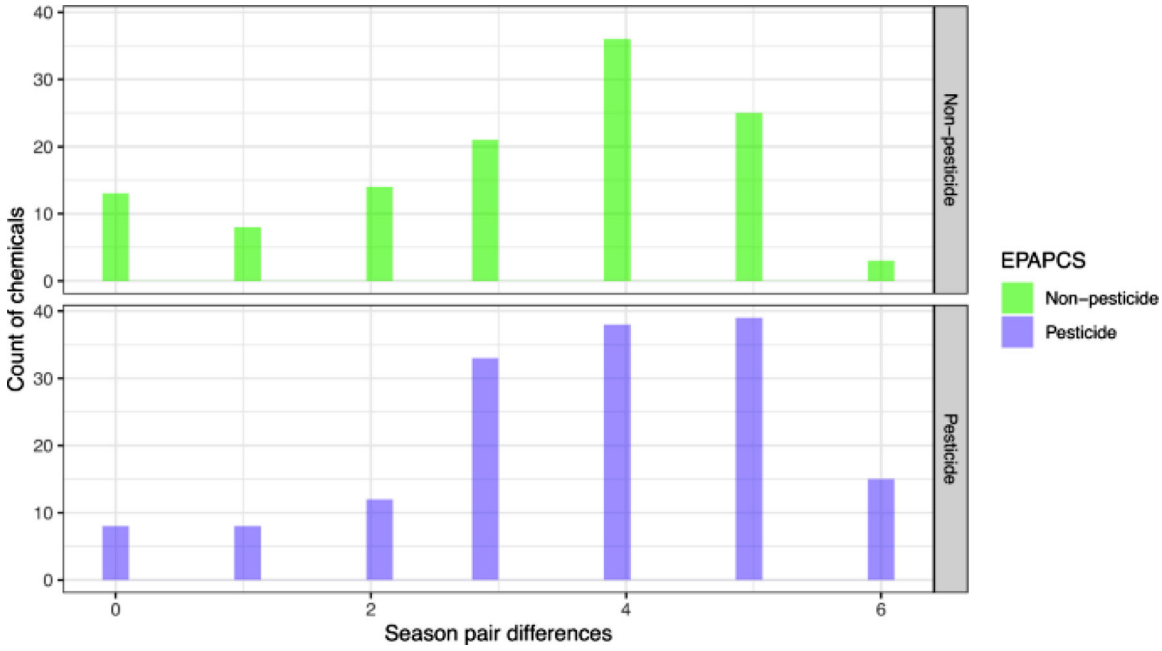

Fig. 3.

The amount of seasonal concentration variation was similar for pesticides and non-pesticides

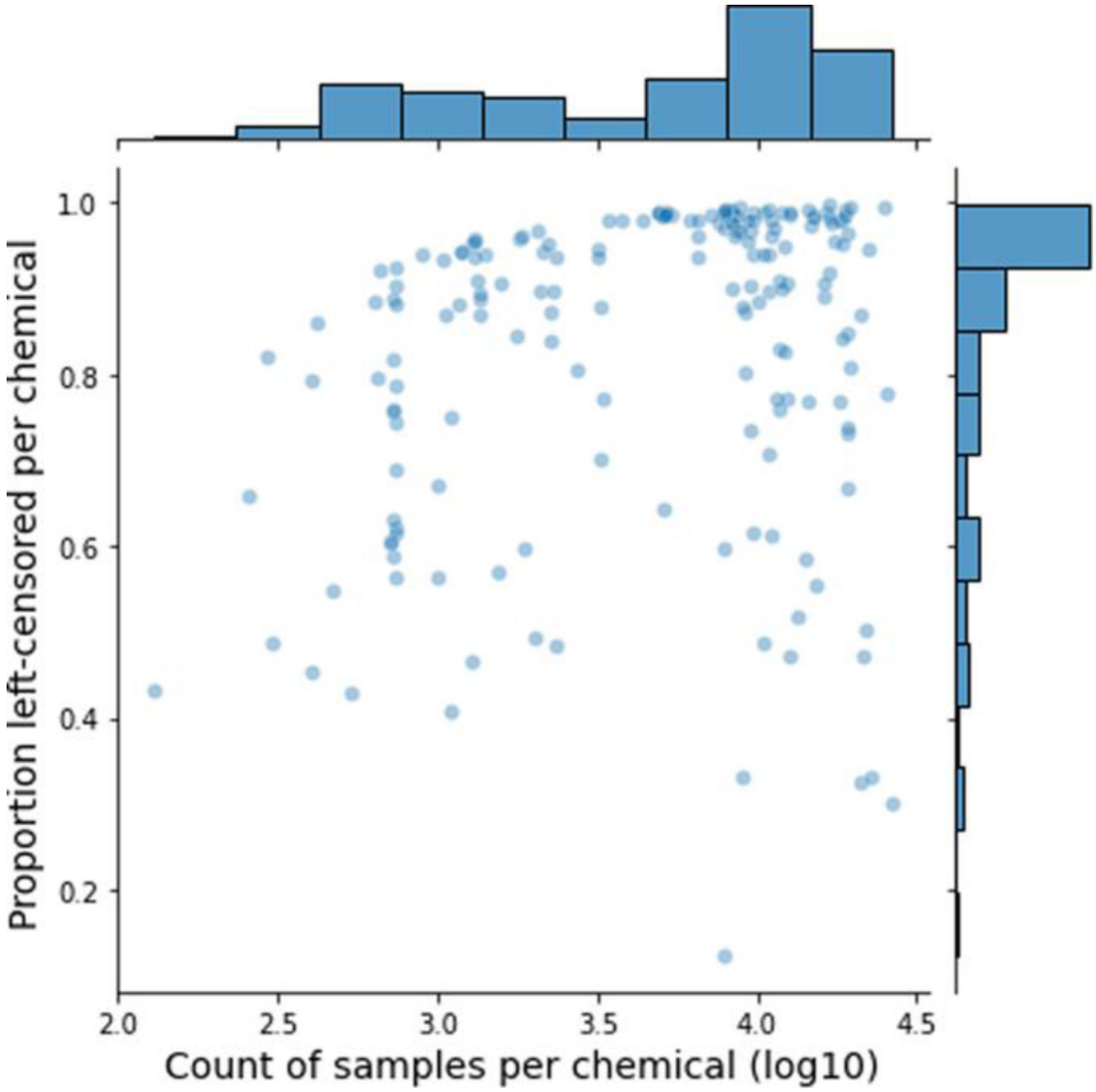

Fig. 4:
Most records are below a limit value.

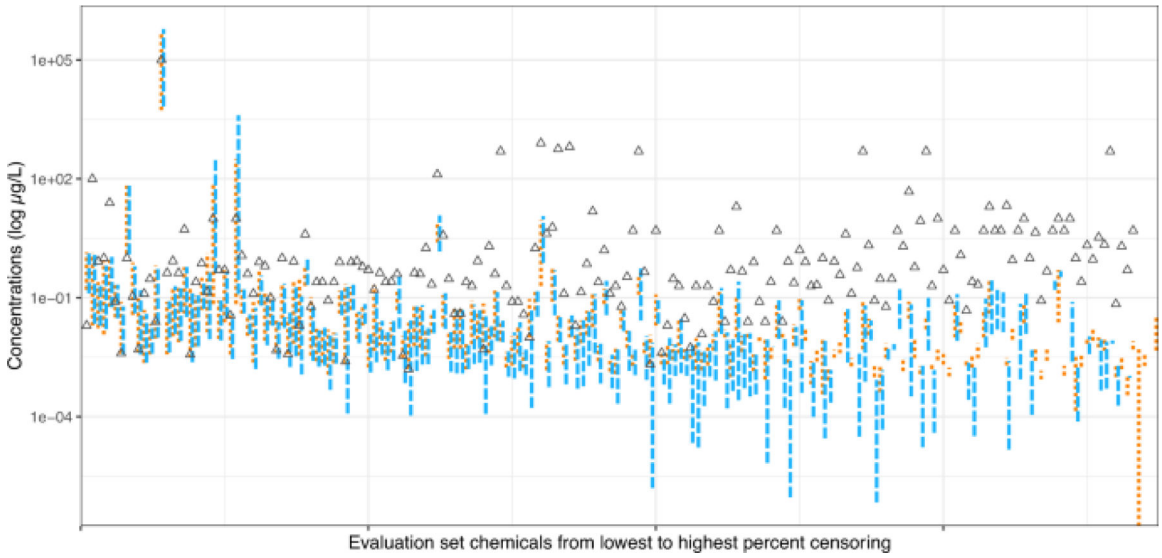

Fig. 5.
Statistical methods agree when data are less censored.

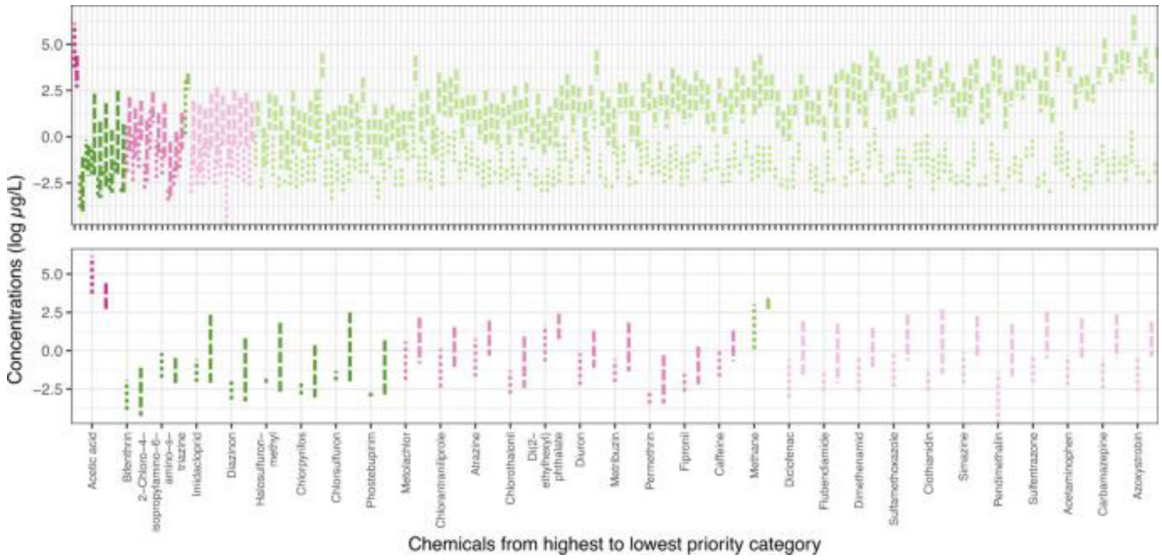

Fig. 6:
Risk prioritization based on bioactivity exposure range overlap

Table 1

Glossary

Term	Definition
Censored data	A measurement where the concentration was below the threshold for detecting the chemical.Also called a “non-detect”.
Technical Min	The lowest possible result detectable using a particular instrument or method (also called a non-detect).
Techinical Quant	the lowest result deemed quantifiable using a particular instrument or method.
Technical Limit	Either a technical min or technical quant, sometimes reported without specifying which.
Reporting Limit	The lowest value reportable based on a laboratory’s certification or other regulatory threshold.
Sample Result	The observed or measured value of the concentration in the water sample, which is only available if the observation is not censored.