Determining Representative National Surface Water Chemical Concentrations for Risk Prioritization

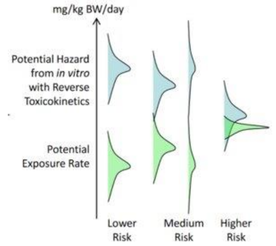
Risa R. Sayre[1,2,3], Marc Serre[2], R. Woodrow Setzer, John Wambaugh[1]

1: U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure; 2: Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill; 3:Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee;

Risa R. Sayre | sayre.risa@epa.gov | ORCiD 0000-0002-6173-8020

1. Introduction

Background: With thousands of chemicals in commerce and the environment, efficient tools are needed to support risk prioritization and evaluation.



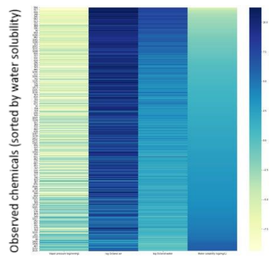
Knowledge gap: Inconsistent data availability for concentrations in surface water to develop exposure estimates.

Proposed solution: Development of an open, reproducible workflow to:

1. Determine representative surface water concentrations for hundreds of organic chemicals in the United States based on already available monitoring data
2. Prioritize organic chemicals based on the relationship between concentration ranges and predicted no-effect concentrations (PNECs) for standard freshwater test species

2. Data overview and curation

The Water Quality Portal (https://www.waterqualitydata.us/portal/) provided millions of concentrations of organic chemicals in surface water sampled from 2008 to 2018 covering broad spatial and physicochemical property ranges.



- 1626 names mapped to chemical structures using EPA's Chemicals Dashboard; 111 names manually curated
- 117 unmapped; 311 names referring to mixtures, ambiguous structures, organometallics manually removed

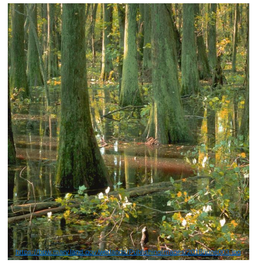
The final dataset contains 1404 unique structures.

Upper right: Sampling sites of observation set represent 2114 of 2270 hydrologic subbasins. Lower left: Chemical property space (log10, calculated using OPERA 2.4) of observation set: vapor pressure (mmHg), octanol:air, octanol:water, water solubility (mg/L).

3. Metadata filtering

Excluded sites:

- Not representative of ambient concentrations (Waste-injection well, sewer, finished water)
- Not surface water (Borehole, atmospheric)
- Not fresh water (Ocean, estuary)



A palustrine wetland

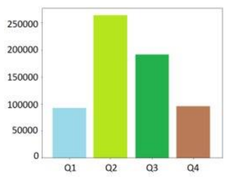
Included sites:

- Surface (some edge cases like palustrine wetland, hyporheic-zone/Ranney well, stormwater)
- Samples labeled as simply "water" without metadata

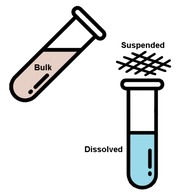
Excluded activities: Not representative of ambient concentrations (blanks, spikes, leachate, initial dilution zone, radiolabeled)

4. Identifying representative subsets

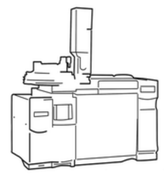
Using two-sample Kolmogorov-Smirnov (KS) tests, we determine whether observed concentrations per chemical are "same" or "different", comparing sets by:



Season
Although more samples were collected in warmer months, observations for about 90% of chemicals were **not significantly different** in magnitude



Sample phase
Determined from three metadata fields. For the 334 chemicals with both bulk and dissolved concentrations, 33% were **significantly different.**

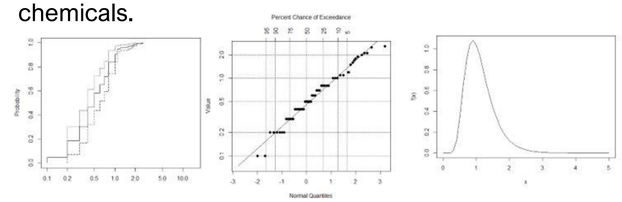


Limit value type
Observations above reporting limits, quantitation limits, and detection limits were **not significantly different**.

5. Estimates of means using censored data

94 chemicals were excluded based on <2 observed values (measurements above the limit value) per sample phase.

We evaluated three different methods of estimation by comparing the confidence interval (CI) around the mean concentration across 20 equal-sized censoring levels from <30% up to >99.5%. In every group, **MLE** had the smallest CI for the greatest number of chemicals.

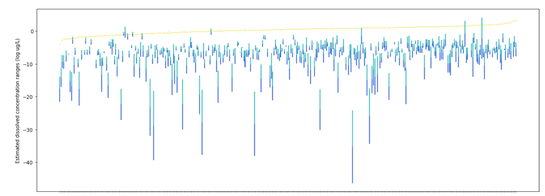


Concentration distributions using (left to right) Kaplan-Meier, robust regression on order (rROS), and Maximum Likellihood (MLE) for single chemical dissolved results with multiple censoring limits and 92.6% censored data.

Chemicals with mean CI >1 µg/L were excluded from further analysis.

7. Prioritization based on ecotoxicity estimates

358 dissolved estimated distributions were compared with PNECs based on the lowest of three TEST-predicted LC50s.



The PNEC was within one standard deviation of the mean for nine chemicals and was not below any range.

8. References

Arnot, Jon A., *et al*. Prioritizing chemicals and data requirements for screening-level exposure and risk assessment. Environmental health perspectives 120.11 (2012): 1565-1570. || Kavlock, Robert J., *et al*. "Accelerating the pace of chemical risk assessment." Chemical research in toxicology 31.5 (2018): 287-290. || Mansouri, K., *et al.* Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J Cheminform* 11, 60 (2019). || Martin, T.M., P. Harten, R. Venkatapathy, S. Das and D.M. Young. (2008). "A Hierarchical Clustering Methodology for the Estimation of Toxicity." *Toxicology Mechanisms and Methods*, 18, 2: 251–266. || Read, E. K., *et al*. (2017). Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research*, *53*(2), 1735–1745. || Wambaugh, John F., et al. "High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals." Environmental science & technology (2014).

This project was supported in part by an appointment to the Internship/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.