**Main comments**

Affordability is mentioned on page 2 (2nd paragraph) but not defined or discussed. What does this term mean within the drinking water context? How does it differ/overlap with the EJ concern about effects on low-income populations? I suggest you weave this thread into the first paragraph since affordability criteria stem from statutory obligations and bring their own set of requirements. This will also broaden the paper in useful ways, I think. I would also suggest moving beyond just measure low income as twice the poverty level to be more inclusive of other possible affordability measures.

The paper references EO 12898 but says nothing about 14096, which should be added and emphasized. 14096 restates the goals of delivering environmental quality “regardless of race, ethnicity, income, etc.” It does not replace minority populations with people of color. Since we recommend against using this aggregate category in reg analysis I recommend de-emphasizing the “people of color” category in favor of using the disaggregated categories found in Table 3. You basically can just set up the aggregate category comparisons as an initial benchmark but really focus on underlying heterogeneity by race and ethnicity.

The literature review on pages 6 -7 does not separate out studies that specifically looked at the sensitivity of results to how service area boundaries are defined vs those that look at disparities in water quality across demographic groups generally. I would organize these into those two bins. Right now, it is confusing because it mixes between them.

For summary stats and perhaps other descriptive statistics it would be useful to better understand the prevalence of zeros and then conditional on having a violation or exceedance the distribution.

I would suggest framing the choice of boundary differently in the paper. You can start with what I think is a relatively uncontroversial statement – or you can pose it as a question – that the county-based service area boundaries is likely the least accurate, but it is unclear how much this matters. You basically set it up as the strawman – it is the simplest to operationalize nationally but we also know it is likely wrong. What is less clear is how much error it introduces and how much does it actually matter for more precisely mapping drinking water quality. Your paper tries to answer this question. Then you can describe the main alternatives – EPIC and Hall & Murray capture almost every water system in the US but they take different approaches to filling in missing data in areas where there is less precise information available. You can then show that the two alternatives to county level designations line up pretty well for areas with better data – at least at the national level, on average – and spend more time exploring the water systems where service area boundaries do not line up across approaches. You have a sensitivity that shows that differences are basically driven by what you assume about Tier 3 systems. I would bring this into the main body of the paper and use it to then explore how much it matters.

I would also structure the paper so you first demonstrate the way these choices affect measures of drinking water quality apart from any discussion of demographics before proceeding to a discussion of the implications for specific demographic groups. For instance, section 5.3 on mapping drinking water quality would come first in the discussion of results as currently structured. Then discussion of relative risks would come next, followed by the regression results.

I understand why you decided to focus on national difference in drinking water quality, but the paper as currently written gives us little insight into how choice of service area boundary can affect differences in drinking water quality across communities in the U.S. This would, however, also make a great addition to the paper.

In particular, I think with some additional work you can unpack at a high level what may drive some of the differences in national level drinking water quality indicators in Tables 2 and 3. While the national averages line up pretty well across EPIC and Hall and Murray for Tiers 1 and 2, does that hold up when you disaggregate by race and ethnicity? Does it hold up across large vs small systems; urban vs. rural? When does the choice of service area boundary appear to matter more/less? And what about for Tier 3?

Another way to approach this is: instead of starting with the service area boundary, start withs the demographics. So, for instance, if you partition the data by income quintile how do differences in assigned service area boundaries affect your conclusions about their water quality? Also, right now you leave the set of bivariate maps until the very end but these could be woven more convincingly into the main body of the paper.

It is awkward to have the USGS and zip code representations of service area boundaries in the same tables and being treated as equivalent to the county, EPIC and Hall and Murray boundaries. These are apples and oranges comparisons. The fact that you find differences in relative risk ratios may be driven by selection since they are available for only a subset of systems. If you want to include them, I have two suggestions:

(1) Create separate tables that only compare the rel. risk ratios across boundary delineations that are limited to the same set of water systems where these data are available. That way we can see what is being driven by differences in boundaries and not by which systems are included. I think I would include these comparisons in an appendix as a sensitivity.

(2) Could you use the zip code and USGS data in a supplementary way? We know the quality of data for Tier 3 systems is poorer. How does what EPIC and Hall and Murray rely on differ (is EPIC just county for example or zip code?) and can we learn anything about these differences by leveraging zip code or USGS data?

You mention on page 6 that CA, TX and VA have the best quality boundaries. This would allow you to weigh directly at least for these three states on boundary accuracy. If you know these are good quality, you can at least compare county to the actual boundaries, and show that EPIC and Hall and Murray (presumably) also match up well. If there is reasonably complete zip code and USGS boundary data then this might also give you a way to assess those approaches (subject to the apples and oranges comment above). You can note the caveat that any quality ranking for these three states doesn’t necessarily apply to other states where we know less about the actual service area boundaries, but at least you can develop insights into when the choice is likely to matter in those cases.

For the regression analyses, can you explain why you preference the EPIC service area boundaries over Hall and Murray? I would run them both and then footnote the H&M ones if they are very similar. It would also be useful to see if using counties producing something quite different. You might consider doing a quantile regression as well to see if the trends are the same for water systems with higher numbers of exceedances or violations. Zeros may dominate here.

The section of the paper that combines information on drinking water quality and demographics with other environmental indicators to see what the relationship is between them. It doesn’t fit well into the paper right now – particularly since it takes you away from comparisons across different ways of delineating service area boundaries. Is the goal to show that existing indicators, while related, do not adequately capture key aspects of drinking water quality? I would argue that you aren’t really quantifying cumulative burden per se, but rather characterizing how environmental indicators may co-occur (or not). I could see an argument for dropping this, but if you want to keep it think hard about the connection. I also recommend that, in addition to recharacterizing it, you perhaps focus on indicators that one would expect to have some relation to/affect water quality and add a discussion of what your prior is with regard to how each of the remaining indicators relate to water quality (Superfund sites would be expected to affect downstream WQ if the contamination leaches into the soil, for instance).

p. 23 onward. The last part of the paper feels unstructured. You are throwing a bunch of maps in, but they aren’t connected to the point of the paper. For instance, for the section labeled nationwide drinking water quality indicators you are presenting county plots. There is no comparison to other boundaries. The bivariate maps seem highly relevant and could be leveraged to much better effect by being moved just before what is currently labeled as section 5.1. The state level bivariate maps also don’t seem to fit well into the current paper. You also offer know explanation for why you focus on these three specific states. Are they chosen at random? Finally, you select health-based violations and EPIC boundaries to show underlying heterogeneity. Why this metric vs some of those with more going on geographically like DP and SDWA violations? Also how different would the map look if you used Hall and Murray instead? Can you more explicitly compare these to the county maps you produced?

**Other Comments**

Abstract

I would delete the second sentence from the abstract. Instead, I recommend you replace it with a statement about what a lack of accurate service boundary data might mean, particularly for analyses where not having an accurate delineation of the community affected may introduce errors in how measures of water quality is assigned geospatially. I would then also more explicitly note how EJ comes in: When these data are combined with information on who lives within these communities, such errors may hider one’s ability to understand whether some population groups generally have poorer water quality than others. This paper examines how several service boundary datasets that differ in accuracy can affect both pieces of this type of exercise: both how it affect measures of drinking water quality and what it means for the types of conclusions one might draw about the distribution of drinking water quality by demographic group.

Introduction

p. 2, 1st paragraph: 9 to 45 million Americans have a drinking water violation? That seems like a wide range. Worthy of a footnote to explain what drives uncertainty about the number of people affected? Or is the 3 – 10% of water systems a range that reflects ups and downs over time?

p. 2 1st paragraph, last sentence: minor clarification suggested in italics – “given the importance of accurately characterizing demographics *within affected communities*,…

p. 2, 2nd paragraph: this is the first use of the term service area boundary products. This term doesn’t sit well with me – maybe because it makes them sound like different formal commercial data products on offer, which they are not. In this specific place, I think you can just reword the sentence to avoid the term. I will make alt wording suggestions later on in my comments, too. Here, you can say: “First, we investigate the extent to which the choice of *how to delineate* service area boundaries may yield different….”

p. 4, last paragraph of the intro: Do these papers characterize themselves as screening analyses or as proximity analyses? I guess I thought they were the latter.

p. 4, last paragraph of intro: Suggest rephrasing of sentence: “By assessing the *EJ* implications of different geospatial boundaries, our contributions are threefold:…”

p. 4, last paragraph of intro (next sentence): suggest deleting the phrase, ”…allowing us to compare different drinking water concerns nationally.”

p. 4, last paragraph of intro (last sentence): suggest shortening and rephrasing the sentence to read: “Lastly, our focus on the extent to which the choice of geographic boundaries affects conclusions *points to the* importance of collecting and *disseminating more accurate* boundaries where possible.”

Background

p. 5, 2nd paragraph: I am not sure why you are citing Wolverton 2023 for E.O 12898. It should be a White House citation. I would think Wolverton 2023 belongs at the end of the next sentence along with Cecot and Hahn 2022.

p.5 and throughout document: you occasionally use the phrase census divisions but that is an actual Census delineated boundary and I don’t think that is what you mean. Perhaps you could use census designations instead?

P. 5, third paragraph: I don’t think this statement is accurate so I suggest deleting the sentence: Traditionally, these analyses assign the demographic information of a county served to the water system itself. Sometimes, I think its been a combination of county and zip codes. Also, you don’t need this sentence since you offer a specific example next, though I would replace “standard practice” with “*was used in EPA’s EJ analysis for the 2020 steam electric rule*” and drop the footnote.

p. 5, last paragraph: I would also recommend deleting the first two sentences here. I would argue an analysis from 25 years ago is not a useful citation. Instead, you can start the paragraph with, “*however,* the EJ analysis for the 2023 steam electric….”

p. 5 last paragraph: *Over time, then*, these analyses have employed…”?

At the very bottom of page 5 you say choices about what approximation of service boundaries to use has proceeded in a vacuum of evidence. This is true and also puts a fine point on why you should weigh in not just on how the choice affects outcomes but also which of them seem most defensible (see above for a few ideas).

I recommend adding footnote 7 directly to the text.

p. 7, very top – this seems like a separate issue with respect to the boundary challenge. Should it be placed elsewhere?

p. 7, You include studies about water rates, which seems like it should be in a footnote?

p. 7, Sometimes you describe a paper it might opine on what difference the boundaries used makes (e.g., Berahzer, et al 2022) but you don’t state the results in those terms. Add?

Data

p. 8, top – sometimes you say population characteristics, sometimes sociodemographic. I would pick one consistent term and use it throughout.

Section 3.1.

p. 8 – you use other state-level datasets aside from the PFAS analytic tools. Mention them here? I would drop reference to EPA RIAs here to make the data description less EPA specific and drop the last two sentences which don’t seem to be germane to the section.

p. 8 : What is the source of data for the LCR ALEs? Is it SDWIS?

p. 8; LCR ALEs – I don’t understand how using an indicator of the number of lead ALEs overcomes the absence of 90th concentration observations for the smallest systems if that is what triggers an ALE. Can you clarify?

p.9 – PFAS: can you explain why you need the state-level sampling data to supplement the UCMR data? This could be a footnote.

p. 9 – PFAS – what proportion of systems have only a single detection?

p. 10 DBP – you do a nice job of explaining why DBP and total coliform are relevant from a health standpoint, but do not include for lead or PFAS. I recommend adding them.

Section 3.2

p. 11 – Service Area Boundary Designations instead of Products? Also you use the term indicators here but I think you used metrics earlier. I prefer measures. I think you can drop reference o other indicators of environmental quality here, as they do not directly hinge on the way the service boundary is determined.

p. 12 – Hall and Murray - here you say CA, CT and NJ have the highest quality boundary information. Earlier you said it was CA, VA and TX – which is it? Is it all five? What is the source of the discrepancy?

p. 12 – Hall and Murray – you discuss what they do for Tier 1 boundaries, but it is less clear how they refine boundaries for the other tiers. Can you explain?

Section 3.3

p. 13 – the first paragraph could use some smoothing out. For instance, you could say, “We use five year 2021 ACS data to assign….” In the first sentence and then eliminate a later sentence altogether. Reference to EJScreen is odd here as well because that is not about pulling population characteristics, right? That is for the other environmental indicators. I would recommend just using ACS directly for those.

p. 13 – I suggest dropping the second paragraph completely an replacing it with a discussion of possible measures of race, ethnicity and low income (with this one also tied to the notion of what is relevant for measuring affordability).

p. 14 – I would move the Mapping Drinking Water Indicators up in the order of this section. Start here.

Results

p. 18 – you state the count of PFAS detected is lower, do you mean the relative risk measure based on counts?

p. 18 - I would restructure Table 3 to include county, EPIC and Hall & Murray boundaries. I am not sure why those results are in an appendix.

Section 5.2

p. 20 – you use Poisson and OLS regressions. Presumably OLS is problematic if there are a lot of zeros? Also wondering about using logs instead for the OLS?

P. 20/21 – can you put the regression results in context. Looking at the summary stats, a 54% increase in the count of health-based violations is relative to 1.3 on average. So, this means areas with Hispanic populations have an average of 2? Likewise for lead ALEs we are talking about 25% more in Black communities but this is relative to an average of 0.45?

p. 21, bottom – you say drinking water quality disparities in Table 3 are statistically indistinguishable from zero. Are these still relative risk ratios? In which case statistically indistinguishable from 1 (no disparity)? I am also confused about the statement with respect to income since Black and income are both statistically significant?

p. 22, Section 5.3. You discuss mapping the EPIC boundaries but what about the others?

Footnote 29 - can you bring some of this discussion into the paper itself?