**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.

*Response:* We remove references to affordability as something we look at in this paper as it would broaden out the scope too much and distract from the main focus of the paper which is the implications of the different ways to characterize water system boundaries.

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.

*Response: We have added a reference to EO 14096 and additional discussion of its relevance to the paper. We agree that the aggregate group of people of color should be de-emphasized in favor of the dis-aggregated categories in Table 3. We have de-emphasized table 3 and replaced most of this discussion of it with description of a series of bar charts that convey a similar comparison but for each demographic group. We have also edited the empirical section to better correspond with these changes to how the results are presented. We retain bivariate maps that look at people of color as an aggregate group because it is more streamlined than presenting a series of bivariate maps for each demographic group and each indicator.*

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.

*Response: Edited and reorganized for clarity. We don’t exactly have two bins, as we were only able to find a single study that did a real comparative analysis, but hopefully the current arrangement is easier to follow.*

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.

*Response: Zeroes as a percentage of PWS have been added to the summary statistics table. We have also added a new table to the appendix (see Table 6) that summarizes average levels for each indicator when excluding all zeroes.*

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.

*Response: We have reframed the choice of boundary product and edited the motivation in the second paragraph of the introduction. We now separate our contribution into one part that compares the county vs. newer improved data and a separate contribution that compares conclusions across all the newer improved data. We have added text throughout the paper that more explicitly compares the worst boundaries (i.e., county) to the better ones (e.g., EPA ORD). We also now reference the appendix table breaking out results for only tier 1 and tier 2 systems in the main body of the text, and we spend more time describing the sources of differential conclusions throughout the paper as elaborated below.*

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.

*Response: We have re-structured the results in this way. We now begin with a discussion of nationwide levels of each drinking water metric according to our summary statistics tables, and then we describe the nationwide maps of these indicators. We then turn to describing the relative risk ratios for each metric, and contextualize these relative risks using the set of bivariate maps. As part of this discussion of relative risk, we incorporate additional description of our selection of particular states and indicators. The description of regressions is now last and significantly shortened, with much of it moved to the appendix.*

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.

*Response: We have added additional discussion of the raw levels of these variables, how they vary across boundary representations, and how they vary across space when employing different boundary products. We now have an explicit comparison of two boundary products that allows us to elaborate on some likely reasons why different service area boundaries can lead to different conclusions. In particular, we highlight three sources of inaccuracies in boundary products. These include:*

1. *Wrong location: It’s not always easy to pin down the right locations, especially given that underlying SDWIS data sometimes provides geospatial data in the wrong state. Different boundary products handle this differently. This type of error could affect conclusions in unpredictable ways depending on the types of systems that are more/less likely to be inaccurately placed.*
2. *Oversized boundary: Boundary products such as counties or zipcodes are often the right position but sweep in large swathes of areas that are not actually served by a system, which can dilute the extent of an EJ disparity if non-served areas are higher income or less likely to be minority populations.*
3. *Improperly intersecting boundaries: The boundary products don’t always de-intersect service areas, due in part to data limitations or to the geospatial extent of the polygon (e.g., counties can serve multiple systems). This type of error can lead to unpredictable bias.*

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?

*Response: We added summary statistics tables for the average levels of each metric by demographic group and according to each boundary. We also would like to add break-out analysis that compares relative risk for small (<10,000 population) and then for large (>=10,000 population) systems. Regarding rurality, given challenges in assigning rural status and how this varies across the boundary locations in each dataset, we leave this for later work. We will plan to leave these potential additions off of the paper that we submit to a journal, as we want to strategically leave some low-hanging fruit for referees to pick up on.*

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.

*Response: We now weave the discussion of bivariate maps more fully into the discussion of relative risks, and we describe differences in levels across demographics before diving into the differences in relative risks. We are not able to partition the data by income quintile using the data previously incorporated using the the areal apportionment code.*

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?

*Response: We feel it’s important to include the zipcode and USGS boundaries for a few reasons. First, the zipcode areas have been proposed by others across EPA as an alternative to counties and also used in multiple RIAs, so providing evidence on them is relevant to Agency considerations. In addition, external researchers using medical data often have only zipcode health records and therefore need to join to DW system information by zipcode, so it's useful to these researchers to show the zipcode comparisons. Second, the extent of completeness across boundaries is one reason for why results can differ, and it seems important to show users that completeness does indeed vary.*

*Separately, it’s worth noting that the service areas may vary a lot in the coverage as a percent of all water systems, but they don’t vary as much as a percent of the total population served. We added a sum of the population served for each boundary to select tables to show that the total population served is not as starkly different as the number of systems included, as the larger systems are captured in most of the datasets, and the large proportion of missing systems serve relatively few people.*

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.

*Response: The EPIC and Hall and Murray boundaries each use the state data where-ever available, and so for high quality states they should be almost identical excepting a small portion of systems that are not included with the state data. The text on page 6 was somewhat unclear, and so we’ve revised it and noted in the paper that the states with high-quality boundaries are actually CT, CA, NM, NJ, WA.*

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.

*Response: We ran the regressions only using the EPIC data for this first pass of the paper because the H&M final boundaries were not yet finalized, but we agree that it would be ideal to include both. There are some differences between the results, and so we’ve decided to only include the H&M-based regressions and move the EPIC regressions to an appendix.*

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).

*Response: We agree that these regressions are somewhat disjoint with the overall goal of the analysis and are fine with re-characterizing this analysis as reflecting co-occurrence rather than cumulative burden. We feel there is some value in retaining the multivariable analysis in case reviewer’s ask how the risks vary when jointly controlling for income, which commenters have raised when presenting, but we have significantly abbreviated the main discussion in the paper, moved most of it to the appendix, and placed all of these tables in the appendix. Since this discussion is now shortened, we do not add any new text about our priors on the correlations.*

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 no 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 DBP 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?

*Response: We’ve re-organized the discussion of the results such that these maps are now described alongside the summary statistics and presented first. In this re-organization, we now emphasize the characterization of drinking water quality as a primary contribution and then follow it with a summary of how DW varies across service area boundary types. We’ve also added national bivariate plots for all indicators. We have woven the discussion of bivariate maps into the explanation of differences in relative risk and explained the rationale for selecting particular states for certain indicators. We have a new discussion comparing the EPIC and ORD boundaries side by side.*

**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.

*Response: We have re-structured the abstract along these lines and added additional references throughout the paper to how use of different service areas affects the perceived average DW quality as well as disparities across systems.*

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?

*Response: The 3-10% reflects ups and downs from 1982-2015. Sentence edited for clarity.*

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

*Response: Revision accepted.*

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….”

*Response: Revision accepted. We agree that the use of the term “product” is not accurate and have removed it throughout the text.*

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.

*Response: We removed the word “screening” from this sentence so that it does not inadvertently mischaracterize those studies or suggest that we are only focused on sensitivity to geospatial techniques in proximity type analysis.*

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

*Response: Revision accepted.*

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

*Response: Revision accepted.*

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.”

*Response: Revision accepted with minor adjustements: “*Lastly, our focus on the extent to which the choice of geographic boundaries affects conclusions of an environmental justice analysis points to the importance of collecting and disseminating more accurate service area boundaries as well as to the value in using the highest-quality boundaries where possible in academic or regulatory analysis.”

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.

*Response: Revision accepted.*

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?

*Response: Revision accepted. Depending on the context, we replace this term with “census geographic units,” “census designations,” or “census block groups.”*

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.

*Response: Revisions partially accepted and additional edits made for clarity and to reference other EPA DW EJ analyses. We retain an explanation for typical EPA approach because the assumption of county demographics is the only method I’ve seen from analyses prior to my arrival at EPA. It was challenging to advocate for use of anything other than counties (including even the state-level boundaries that were available) because the status quo and precedent of counties was so entrenched. The 2023 steam electric ELG’s use of county + zipcode was actually a concession that OW made in lieu of using available state-level boundaries that I had suggested. Even the most advanced DW EJ analysis, the PFAS NPDWR, still uses county demographics for the entire SafeWater module of the EJ chapter.*

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….”

*Response: This is just a transition sentence. We feel it’s an innocuous anecdote that helps to set up and contextualize the advancements EPA has made in EJ analyses.*

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).

*Response: We prefer not to weigh in on which boundaries to use for a few reasons. Most importantly, ORD let us know that they would like to write a paper that explicitly compares accuracy of boundaries, and so we do not want to step on their toes. Related to this point, we don’t actually observe the accuracy of all boundaries except anecdotally, and so making claims on which boundaries to use would be stepping outside of the evidence we generate. Third, the choice of boundary is partly a policy call based on the relative importance the analyst perceives in having the most accurate analysis vs. the time cost of conducting more advanced spatial analysis. We want to avoid making specific policy recommendations.*

I recommend adding footnote 7 directly to the text.

*Response: Revisions accepted.*

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

*Response:*

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

*Response:*

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?

*Response: We’ve revised the discussion of this paper.*

Data

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

*Response: We prefer to retain these synonymous descriptors so that the language is less monotonous.*

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.

*Response: Revisions accepted.*

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

*Response: Yes, it is SDWIS. Edits made for clarity.*

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?

*Response: We observe ALEs and their associated 90th percentile sample value for all systems irrespective of their size. However, we do not observe the 90th percentile sample values for small systems that do not have an exceedance in a particular compliance period. We have added clarifying text to the paper to note that ALEs are presumably always known but 90th percentiles are not necessarily observed if there is no ALE.*

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

*Response: Footnote added.*

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

*Response: Of the nearly 15,000 systems that have at least some PFAS samples, 73.4% have no detections. The highest number of unique PFAS detected is 18, and 99% of all systems sampled have eight of fewer detections. We’ve included a histogram below and have added the percent of 0s across all indicators to the paper’s summary statistics table.*

**

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.

*Response: We’ve added a discussion of health impacts with respect to PFAS and lead.*

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.

*Response: We systematically replaced all references to service area “products” with either service area “representations” or service area “types.” We also replaced most uses of the term indicator with either measure or metric. In some cases, we vary the use of measure, metric, or indicator to avoid being overly repetitive.*

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?

*Response: We have revised the text to clarify the specific states that have high-quality boundaries. The discussion of one paper’s use of VA boundaries may have suggested they were high-quality, but we do not think that is the case.*

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?

*Response: Hall & Murray’s approach does not separately contend with tier 2 and tier 3 systems, which are a designation unique to the EPIC boundaries. Hall & Murray instead use several techniques to match systems to locations or predict locations. We did not have a particularly firm basis for describing these methods before seeing their documentation, but we have now incorporated significantly more discussion of their methods.*

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.

*Response: Revisions accepted and others made throughout this section to improve readability. We pull ACS statistics via the EJSCREENbatch R program in a function based on TidyCensus, but we also use some EJSCREEN indicators and variables in the paper (e.g., % people of color, environmental indicators).*

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).

*Response: Revisions accepted*

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

*Response: Revisions accepted.*

Results

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

*Response: Yes, although this would also imply that the count of PFAS detected is lower on average for this group. We’ve clarified this in the text.*

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.

*Response: We have combined all results into one table, although we now use bar charts to convey differences across boundary types and relegate the tabular formats to the 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?

*Response: We chose to use Poisson regressions because for outcomes that are counts, but these could be re-run as OLS and included in the appendix if necessary. We did not use logged versions of any outcome because of the presence of zeroes. If necessary, we could incorporate the inverse hyperbolic sine of these outcomes and present those results in an appendix. Since we’re relegating all of this work to the appendix, it does not seem like a particularly important change.*

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?

*Response: We have added more contextual descriptions to these outcomes.*

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?

*Response: These outcomes are PWS-level water quality metrics rather than the relative risk ratios. We have revised the text to clarify that Black and income are statistically significant in certain regressions.*

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

*Response: We have added maps that are based on the Hall & Murray boundaries and explicitly compare the appearance of these to those based on the EPIC maps.*

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

*Response: Edits made.*