Spatial Patterns in Water Quality Portal Data:
Identifying and Addressing Gaps in Water Quality Monitoring & Reporting

21 April 2022

Blair Johnson & Sophia Bryson || Authors
Martin Doyle, PhD || Advisor
Ashley Ward, PhD – Internet of Water || Client

Master's project submitted in partial fulfillment of the
requirements for the Master of Environmental Management degree in
the Nicholas School of the Environment of
Duke University

EXECUTIVE SUMMARY

The Water Quality Portal (WQP) is a collaborative service that integrates publicly available water quality data from three federal databases. It contains monitoring data from 1.5 million sites across all 50 US states, provided by over 400 agencies and organizations. Accordingly, the WQP offers a relatively comprehensive picture of the water quality monitoring activities in the United States. Robust and accessible water quality data can offer both insight and foresight that allow for efficient and timely responses by governments, service providers, and individuals to water quality concerns or threats, in service of the health and wellbeing of human and aquatic life. However, barriers can prevent water quality data providers from performing data collection, reporting, and integration, leaving gaps in the availability and usability of water data. In this project, data from the WQP were analyzed to determine spatial patterns in monitoring coverage at the subwatershed scale. These patterns were integrated with population, race, ethnicity, and income data from the American Community Survey (ACS) to estimate population demographics and characteristics at the subwatershed scale and to determine what communities may be underserved by water quality monitoring.

Subwatersheds were assumed to be monitored if they contained at least one monitoring site for which data had been uploaded to the Water Quality Portal in the past 2 calendar years, or were immediately downstream from a subwatershed containing such a site. Populations were considered to be served by monitoring if the subwatershed in which they live was monitored. During the period of record, there were 290,000 unique instances of water quality monitoring reported for 57,000 distinct sites across all 50 states and the District of Columbia, reporting on over 1000 unique parameters, pertaining to basic water quality as well as to the regulatory standards of the Safe Drinking Water Act and the Clean Water Act. 19% of the 95,000 subwatersheds in the United States were classified as monitored, and 57% of the population lives within these monitored subwatersheds.

The characteristics of populations living within each monitored and unmonitored subwatersheds were compared to assess if the populations served and unserved by monitoring are distinct in their racial, ethnic, and economic characteristics. White, Native American, and Non-Hispanic/Latino individuals were marginally more likely to live in unmonitored subwatersheds than other racial and ethnic groups, and unmonitored subwatersheds had a moderately lower median household income than monitored subwatersheds. The population factor with the strongest connection to monitoring was population density: monitored subwatersheds were an order of magnitude higher in their population density (14.3 persons/sq. km) than unmonitored subwatersheds (1.2 persons/sq. km).

Trends in monitoring, coverage, and data providers were also analyzed by state and by EPA region. 418 unique organizations and agencies provided data during the period of record. States and regions differed in the number and type of monitoring sites and data providers, but state agencies were consistently the foremost provider of data across all spatial scales, with 73% of all monitoring sites during the period of record provided by state agencies. Monitoring coverage also differed by state and region, ranging from 5% to 97% of state population and 31% to 73% of EPA region population living within monitored subwatersheds.

Recommendations for policy and practice that aim to increase both data collection and data reporting were developed to increase the comprehensiveness and accessibility of water quality monitoring data. These recommendations target constraints on incentives and capacity of would-be data providers at the

local, state, and federal levels to increase data collection and integration, and include suggestions for collaborative data management, increased access to and use of available funding sources, and partnerships between states and other organizations and agencies.

➢ **Water quality monitoring** can be improved through providing more guidance on Clean Water Act Section 106 funding to water quality monitoring organizations to ensure organizations have the tools needed to apply for these grants. This can be achieved through state environmental agencies incorporating grant funding application guidance in their water quality monitoring strategies. State agencies can also build partnerships with local agencies and NGOs in order to enhance the level of monitoring through dispersed efforts while ensuring the benefits of water quality monitoring are maximized.

➢ **Water quality reporting** can be improved through improving accessibility to water data management systems such as the Ambient Water Quality Management System (AWQMS) through grant funding and collaboration. Cross-organizational collaboration can improve efforts to upload data by allowing organizations to share data management resources and costs associated with integrating data into the WQX/WQP on a regular basis. In addition, more U.S. states should adapt open water data initiatives that center around improving water quality reporting and integration through developing water data platforms that are compatible with the Water Quality Exchange (WQX).

These findings and recommendations were compiled into a dashboard that serves as an information delivery tool to allow decision makers to investigate patterns in the occurrence of reported water quality monitoring, accessible at:

https://www.arcgis.com/apps/dashboards/0dbe111a2c1542a4a1ff01387b037d13.

Users can explore the and visualize the distribution and coverage of water quality monitoring by watershed, state, and EPA region and can access site-level data about the frequency and nature of monitoring. The code supporting the analysis and dashboard is publicly available through a GitHub repository and can be used to update the analysis as new data are added to the WQP and new ACS estimates are released.

This project was prepared for the Internet of Water, in service of their goal to leverage data to enable better water management.

CONTENTS

INTRODUCTION

Water data can be a crucial tool for allowing individuals, service providers, and government organizations to make informed decisions to protect and promote human health and wellbeing. Accurate, complete, and timely data can provide an empirical basis and support for environmental and health decision making, allowing for insight into the quality and safety of water and into the sources of contamination or degradation that may be causing water quality issues. Water quality data can offer both insight and foresight that allow for efficient and timely responses by governments, service providers, and individuals to water quality concerns or threats. When data about water quality are available, accessible, and understandable to all, the data can provide a powerful means for empowering individuals and organizations with information that ought to inform how decisions around water are made. As water resources are threatened by aging infrastructure, climate change, and over extraction of groundwater resources, access to transparent water data becomes all the more important to guide decision-making and ensure that governmental and non-governmental institutions and agencies can support human and environmental health.

The availability and accessibility of water quality data to decision makers is critical for the benefits offered by water quality monitoring to be realized. The Water Quality Portal (WQP) is a tool that seeks to address, in part, the problem of inaccessible and fragmented data by providing water quality data from a variety of governmental agencies, tribal authorities, nongovernmental organizations, and other data providers in a single discoverable, standardized, and accessible location and format. The WQP offers current and historical water data collected by over 400 state, federal, tribal, and local agencies for more than 1.5 million sites across the nation, all accessible and available for download through an online public portal. Due to the WQP's comprehensive nature, in addition to the positive benefits provided by accessible data, it also offers a valuable opportunity to examine where data are not available by examining gaps in records. When and where water quality is not monitored or is not reported, or where that monitoring and reporting are inadequate, the benefits and protections provided by water quality data monitoring will be correspondingly absent or reduced.

In previous literature, Ward *et al.* highlight the gaps that can exist both within water quality monitoring and between data monitoring/reporting and decision making, creating a landscape that is "data-rich but information-poor"[1]. Kumpel *et al.* detail that this disconnect can occur even when regulatory monitoring programs are in place[2]. This failure in effective monitoring can and does occur at various steps in the process, from failures to collect data due to lack of resources or institutional support, to failures to properly or consistently report data, to a failure to use or implement information due to a lack of awareness or accessibility. However, where data do exist, they can provide an effective tool in service of human health, as Li *et al.* highlight in their discussion of spatial analysis of water quality data's connection to public health[3]. Their analysis also adds some nuance to what constitutes a gap in monitoring data, expanding the suite of tools that may help address insufficient or absent data by creating a framework for spatial interpolation of data. While moving from raw data to informed decision making is a problem far

[1] Ward, R. C., Loftis, J. C. & McBride, G. B. (1986). The "data-rich but information-poor" syndrome in water quality monitoring. *Environ. Manag*. *10,* 291–297.
[2] Kumpel, E., MacLeod, C., Stuart, K. et al. (2020). From data to decisions: understanding information flows within regulatory water quality monitoring programs. *npj Clean Water, 38*, 1-11.
[3] Li, H., Smith, C. D., Wang, L., Li, Z., Xiong, C., Zhang, R. (2019). Combining spatial analysis and a drinking water quality index to evaluate monitoring data. *Int. J. Environ. Res. Public Health, 16,* 357-366.

larger than this project, identifying patterns in data availability and in the policy landscape that contributes to those patterns can be a first step in more effectively generating, distributing, and using data. Understanding who is providing the data is also crucial for understanding, and recommending on, the policy landscape. Josset *et al.* note that state agencies are largely responsible for the enforcement and monitoring of both state and national regulations, with federal agencies disseminating the state-level data sets[4]. By identifying what agencies and organizations are (or are not) providing data, policy initiatives can be targeted at providing support and resources where most needed to support ongoing, or develop new, monitoring efforts.

Accordingly, this project sought to identify gaps, both spatial and temporal, in data provided through the water quality portal. Because the data available through the WQP offer a fairly comprehensive overview of water quality monitoring activities in the United States, identifying gaps in the data can suggest where increased monitoring activity could be fruitfully implemented to help realize the benefits provided by regular water quality monitoring. However, because the WQP is not entirely comprehensive, an analysis of the data it provides may also indicate areas where increased data reporting or integration is needed to improve the discoverability and usability of data generated by existent water quality monitoring activities.

The significance of these gaps was considered primarily in the context of human populations and public health, assessing the size and characteristics of populations with inadequate or non-existent water quality data using data available from the US Census Bureau. The number and type of data providers was also considered in order to better understand the sources of water quality monitoring data. Observation on the sources and locations for which water quality monitoring data are reported provided forms the basis for a policy analysis, conducted to assess the causes of insufficient monitoring and reporting. The policy analysis considers questions of each substance, implementation, and enforcement. From this background, potential solutions arising from changes to policy, process, and partnerships in pursuit of more robust and consistent monitoring and reporting are proposed and discussed.

The analyses were conducted at the subwatershed (HUC12) scale. The choice to use hydrologic boundaries, rather than political or administrative boundaries, as the primary unit of analysis was motivated not only by the correspondence of hydrologic basins with the spatial validity of monitoring data, but also by a desire to encourage decision makers to structure their thinking in terms of these units. While political and administrative boundaries are used in some of the display and analyses, these are only as aggregations of the analyses conducted at the subwatershed scale.

The data, analyses, observations, and recommendations presented here are also provided as an interactive online dashboard. This dashboard is oriented towards information delivery to allow decision makers to better understand and engage with the spatial and temporal patterns of water quality reporting, the tangible human impacts that these patterns may have, and potential options for altering these patterns for the better.

This project was completed for the Internet of Water in service of their goal to leverage data to enable better water management. Assessing and improving the data discoverable and accessible in the Water Quality Portal, and providing a tool that allows for decision makers to better understand patterns in these data, will further their goal of providing better data in pursuit of better water management.

[4] Josset, L., Allaire, M., Hayek, C., Rising, J., Thomas, C., Lall, U. (2019). The U.S. Water Data Gap-A Survey of State-Level Water Data Platforms to Inform the Development of a National Water Portal. *Earth's Future. 7,* 433-449.

DATA SOURCES

WATER QUALITY PORTAL

The Water Quality Portal[5] (WQP) represents a collaborative project by the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and the National Water Quality Monitoring Council (NWQMC) to provide an integrated source for water quality data. The WQP includes publicly available water quality data from the USGS National Water Information System (NWIS), the EPA Water Quality Exchange (WQX), and the USDA Sustaining the Earth's Watersheds – Agricultural Research Database System (STEWARDS), with a common format and access portal[6]. These various data repositories represent data collected and submitted by federal, tribal, state, and local governments, as well as academic, non-profit, and private institutions. The WQX serves as a standardized data protocol for external data providers to submit data to the WQP for public access. Metadata and data are provided for each monitoring site, indicating the organization conducting the monitoring, sampling methods, sampling results, and geographic coordinates.

Because of its integrated nature, the WQP offers a uniquely comprehensive view of water quality monitoring in the United States, both allowing for insights into water quality itself and into broader trends in monitoring and reporting of data pertaining to the same. It is primarily with this latter usage that the current project is concerned. Accordingly, data in the dashboard present summarized and aggregated syntheses of the data available through the portal. Users of the dashboard are encouraged to access data directly from the WQP to supplement the insights provided by these summaries.

AMERICAN COMMUNITY SURVEY

The American Community Survey[7] (ACS) is a product of the United States Census Bureau that provides detailed demographic information for the United States. Unlike the Census, which is conducted on a decennial basis and provides a comprehensive snapshot of particular demographic characteristics of the US at the time of its administration, the ACS is conducted on a rolling basis on a representative sample of American households to gather detailed insights into longer-term trends. ACS insights are currently provided based on 1- and 5-year estimates, with finer geographic resolution provided for the longer-term estimates.

The 2019 5-year ACS (2014-2019) was selected for this project, as it was the most recent 5-year estimate available at the time of analysis[8], and the 5-year timescale provides more stable insights and allows for finer geographic resolution than can be obtained with 1-year estimates. This allows for data on race and income to be obtained at the block group level, and data on ethnicity to be obtained at the census tract level. As new ACS products are released by the Census Bureau, the initial pull script can be modified and rerun to access the most up-to-date demographic data.

[5] https://www.waterqualitydata.us/
[6] https://www.waterqualitydata.us/wqp_description/
[7] https://www.census.gov/programs-surveys/acs
[8] The 2016-2020 ACS 5-year data products were made publicly available on March 17, 2022. This release was delayed due to methodological revisions in response to the COVID-19 pandemic, and was prohibitively late for use in the present analyses.

In addition to the block group and tract level data on race, ethnicity, and income, the geographic boundaries of states and counties were obtained from the Census Bureau's Topologically Integrated Geographic Encoding and Referencing[9] (TIGER) database for use in both analysis and cartography.

National Hydrography Dataset Plus Version 2

The National Hydrography Dataset Plus Version 2[10] (NHDPlusV2) is a product of the EPA Office of Water with the assistance of the USGS. It provides geospatial data on the hydrographic features of the United States at a variety of scales. The NHDPlusV2 includes the National Watershed Boundary Dataset (WBD), which contains multi-scalar geospatial products delineating drainage basins in the US.

Drainage basins can be presented at a variety of geographic resolutions, uniquely identified by hydrologic unit codes, or HUCs, to which additional digits are added to specify increasingly fine resolution. In the system developed by the United States Geological Survey (USGS), the 12-digit hydrologic unit code (HUC12) corresponds to the smallest geographical area for which drainage basins are delineated. These catchments are identified by a unique string of 12 digits that corresponds to a specific drainage basin, and, on average, have an area of 40 square miles. These subwatersheds (henceforth used interchangeably with HUC12) are contained in, and were accessed through, the WBD contained in the NHDPlusV2. These subwatersheds were used as the primary unit of analysis when considering the distribution and coverage of water quality monitoring data.

In addition to the HUC12 data used in our analysis, NHDPlusV2 was also used to obtain stationary map layers of surface waters (rivers, lakes, oceans, etc.) for cartographic purposes.

[9] https://tigerweb.geo.census.gov/
[10] https://nhdplus.com/NHDPlus/

DATA ANALYSIS

All data access and wrangling was accomplished through R, with full code available at the following link:

https://github.com/sab159/MP_SpatialPatternsWQPData.

These scripts access and use data from the sources introduced above to determine the level of monitoring provided by data available through the Water Quality Portal at the subwatershed scale.

Monitoring sites were considered to provide spatial monitoring for the subwatershed in which they are located and for any subwatershed immediately downstream of that subwatershed. Temporally, monitoring sites were considered valid for 2 calendar years. Any subwatershed containing at least one monitoring instance within the past 2 years was considered to be 'monitored'.

The population affected by monitoring activity, or the lack thereof, was based on the population characteristics of census geographies overlapping each subwatershed in proportion to their areal overlap.

The analyses resulted in 2 spatial layer outputs: a point layer showing the locations of all monitoring sites during the period of record with summarized information about the data provider, monitoring frequency, and number and nature of parameters measures; and a polygon layer showing the subwatersheds of the United States with their attributed level of monitoring and population composition.

This subwatershed layer connected the level of monitoring as reported through the Water Quality Portal to the human populations (un)served by those monitoring activities, allowing for the relationship between monitoring and various demographic characteristics to be assessed.

A full explanation and discussion of analysis methods and assumptions can be found in the Appendix.

DASHBOARD CREATION + USAGE

The WQP sites, provider data, and race and income data are integrated into an ArcGIS Dashboard, allowing users to explore and analyze trends in water quality monitoring by watershed, state, and EPA region. The dashboard also facilitates exploration of the relationship between monitoring and race, income, and population density, as well as the number and nature of data providers, at each of these spatial scales. In all cases, the primary unit of analysis was the HUC12 subwatershed, but states and EPA regions can be used to select which monitoring sites and subwatersheds are included in the analyses. The dashboard can be viewed at the following link:

https://www.arcgis.com/apps/dashboards/0dbe111a2c1542a4a1ff01387b037d13

The dashboard is intended as information delivery tool as a form of decision support, presenting layered information allowing users to gradually develop a more thorough understanding of the data and patterns as appropriate to their purposes. The amount of data and analysis presented in the dashboard is intentionally limited in an effort to avoid each overwhelming users and unnecessarily replicating information that can be more appropriately accessed through the Water Quality Portal. Links are provided on the dashboard to the Water Quality Portal, to the source code GitHub repository, and to the webpage for client Internet of Water.

The dashboard (shown in Figure 1) provides an overview of the purposes and methods of this project, presents the data interactively and visually, and offers preliminary observations and recommendations related to the spatial distribution of WQP data.

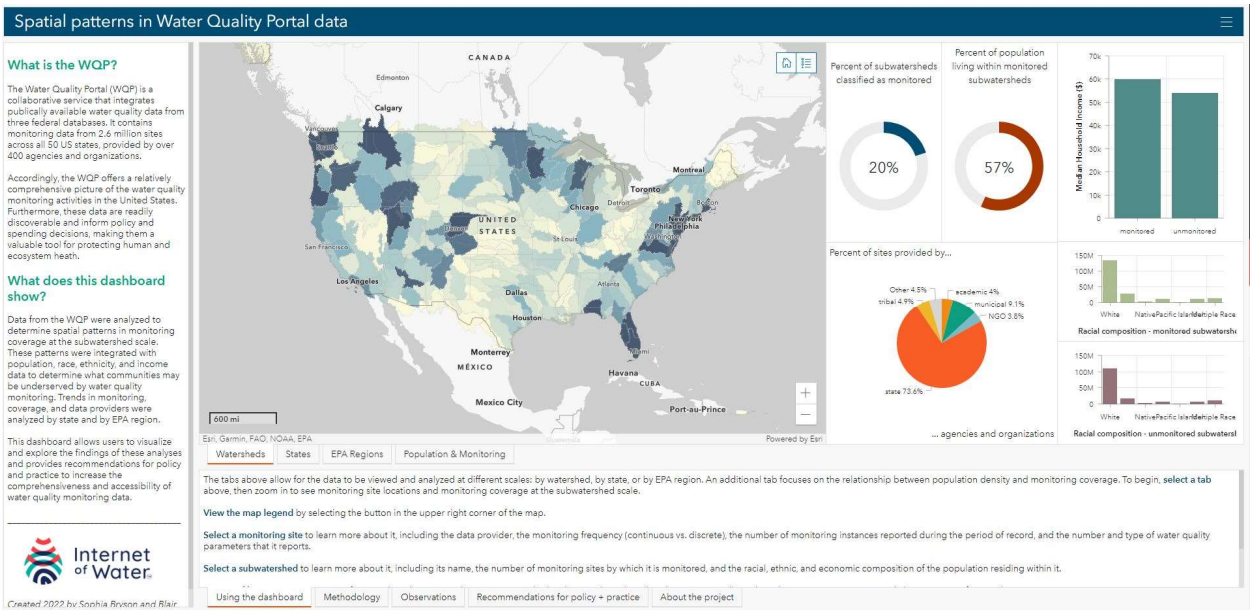

Figure 1 - Opening screen of the dashboard.

The maps on the dashboard show the locations of all monitoring sites from the Water Quality Portal from the period of record as described in the methods. By selecting a monitoring site point (see Figure 2), dashboard users can explore the site's name, provider and provider classification, and whether it monitors surface or groundwater. Information about the site's status as continuous or discrete, and the number of

water quality parameters measured at a site pertaining to general water quality, Clean Water Act designations, and Safe Drinking Water Act standards can also be seen.

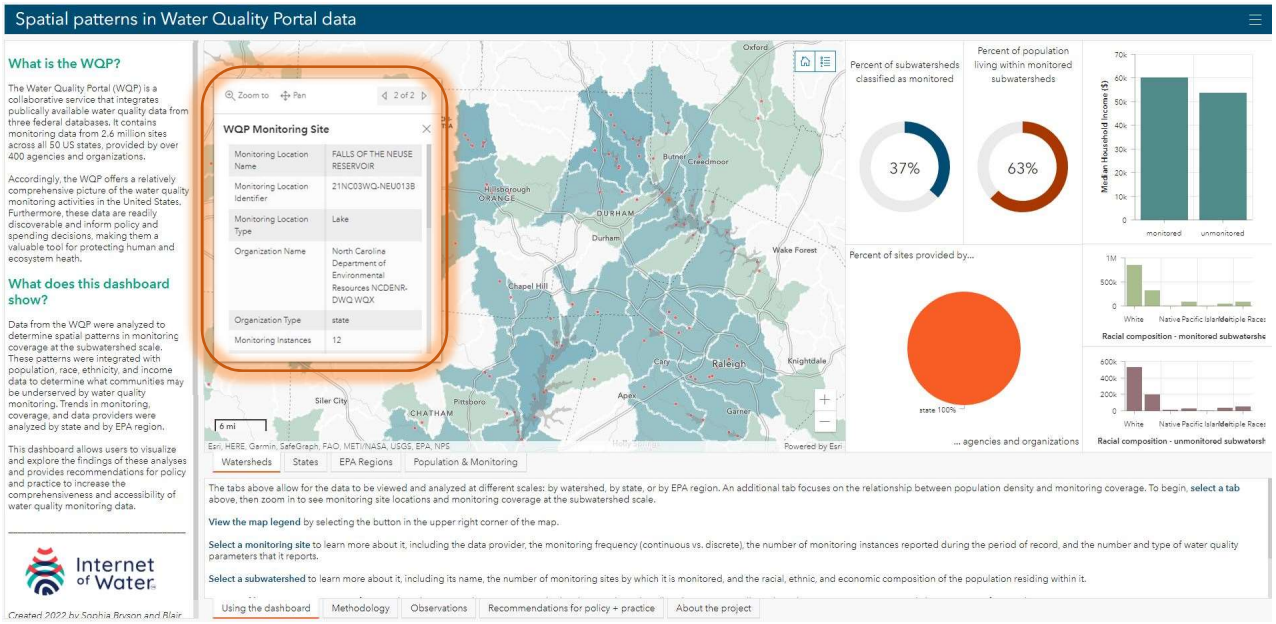

Figure 2 - Dashboard pop-up providing monitoring site information.

The maps on the dashboard also show the level of monitoring of each subwatershed in the United States as described above. By selecting a subwatershed (see Figure 3), dashboard users can see whether a subwatershed is monitored by site data available through the WQP and the number of sites within, upstream, and attributed to the subwatershed. The name, the 12-digit hydrologic unit code (HUC12), and the area of the subwatershed in square kilometers are also shown, as are the population and demographic information interpolated to the subwatershed from census data.

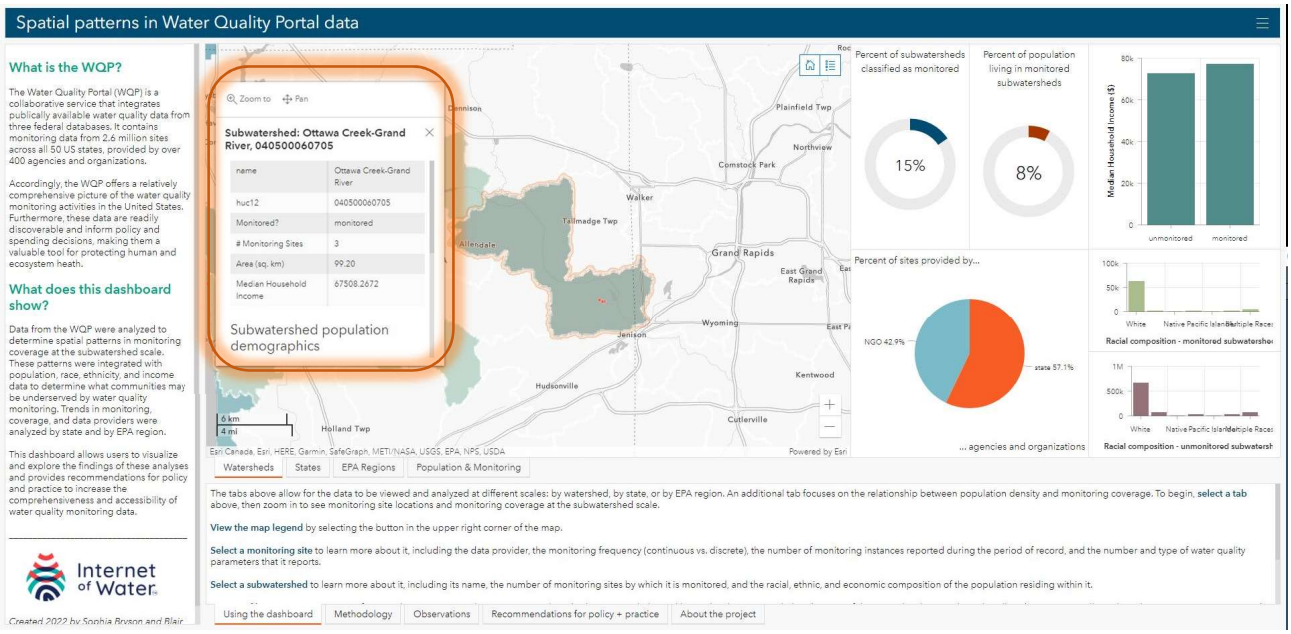

Figure 3 - Dashboard pop-up providing subwatershed information.

The different map tabs on the dashboard allow for users to then explore the extent, source, and coverage of monitoring at the subwatersheds scale within the jurisdictional boundaries of either US states or regions of the US Environmental Protection Agency.

The charts and figures on the dashboard (Figure 4) show the percent of subwatersheds and percent of population that are served by water quality monitoring within the selected jurisdiction (or, if no jurisdiction is selected, the current map extent). Additionally, there are figures showing the proportion of sites provided by different categories of data provider and the income and racial characteristics of monitored and unmonitored subwatersheds.

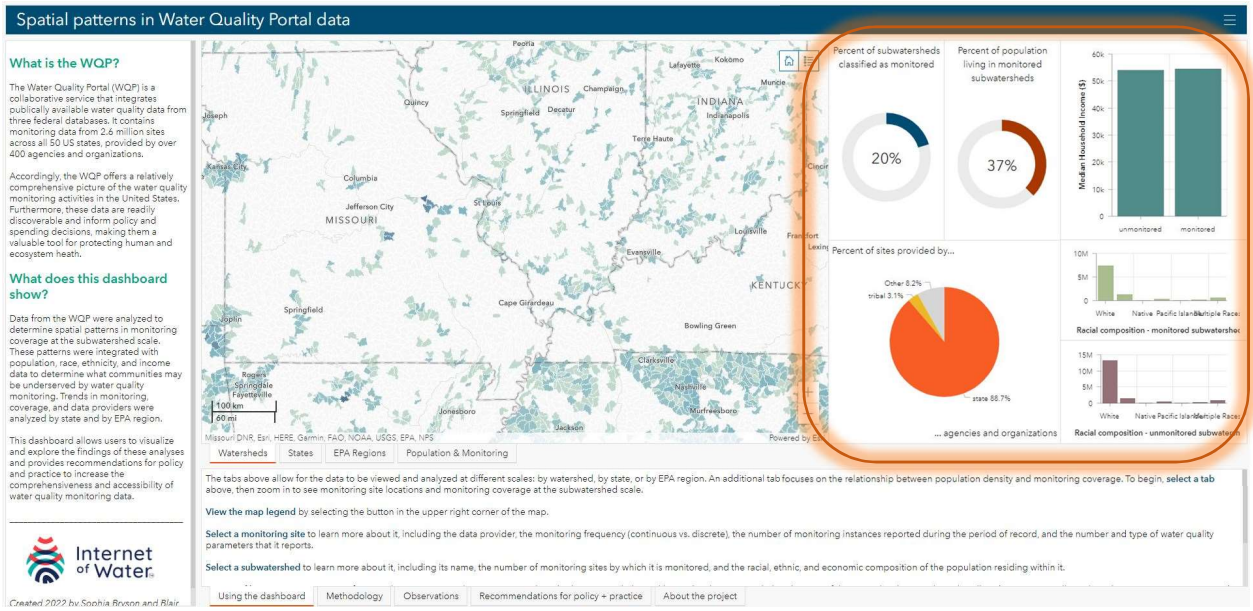

Figure 4 – Dashboard charts showing characteristics of user selection.

Because the dashboard is based on the hosted feature layers updated by the data update script (discussed in the Appendix), the functional content of the dashboard will update automatically in response to updates pushed to the associated feature layers. There are several layers created solely for visualization and not used in any analysis are the exception to this: the layers showing the abundance of monitoring sites at the HUC4 and HUC8 scales are static due to the nature of their creation. These layers are only used to provide a clean visualization before a user zooms into an area of interest on the map, and can be manually recreated if significant changes in the spatial distribution of water quality monitoring have occurred. This can be accomplished by using the 'summarize within' capabilities of AGOL on the HUC12 data with HUC4 and HUC8 layers available in ArcGIS Online's Living Atlas.

ANALYSIS OF FINDINGS

Unless otherwise noted, the following numbers and analyses represent a record period from the beginning of calendar year 2020 through the date of last update, 26 March 2022 (approximately the past 2 years). For that period of record, approximately 290,000 unique instances of water quality monitoring are provided for over 57,000 distinct sites. These data were provided by 418 distinct agencies, organizations, and entities from all 50 states, as well as the District of Columbia. Over 1,000 unique metrics or parameters of water quality were measured across all monitoring instances, pertaining to basic water quality as well as to the regulatory standards of the Safe Drinking Water Act and the Clean Water Act. Of the approximately 57,000 monitoring sites, 87% sample surface water and 11% sample groundwater, with the remaining 2% sampling other media or forms of water. 2% of the sites provide continuous monitoring data, with the remaining 98% represent discrete monitoring activities. 3% of all monitoring instances are from sites providing continuous monitoring.

Of the nearly 95,000 subwatersheds (HUC12) in the US, 17% contain at least one monitoring site, and 19% are considered monitored by our classification (that is, they, or a subwatershed from which they are immediately downstream, contain at least one monitoring site during the period of record). When all monitoring sites from the WQP are included without restriction to the period of record, the proportion of monitored subwatersheds rises to 85%. This indicates that analyses of what geographic areas and populations are served by monitoring data are sensitive to the temporal scope of the considered data. Further inquiry could fruitfully assess the appropriateness of the two-year period of record over which data were considered to provide monitoring coverage in the current analysis.

The demographic data used to assess the human populations affected by the presence or absence of water quality monitoring represent over 335 million individuals (for racial and ethnic considerations) and nearly 125 million households (for income and economic considerations). 57% of the population lives within monitored subwatersheds.

The following sections further examine the nature and distribution of the water quality monitoring activities represented by the WQP data, as well as the providers of those data and the demographics of the populations living in the areas served by the monitoring. The spatial distribution of these characteristics is examined nationally and for each state and EPA regional scales.

MONITORING SITE CHARACTERISTICS & DISTRIBUTION

The monitoring data considered in this analysis represented a diversity of states, regions, location types, provider organizations, and water quality parameters. Monitoring sites varied in the number of unique monitoring events which they represented, ranging from 1 to 24 during the period of record (Figure 5). The majority of sites provided data for only one or two unique monitoring events.

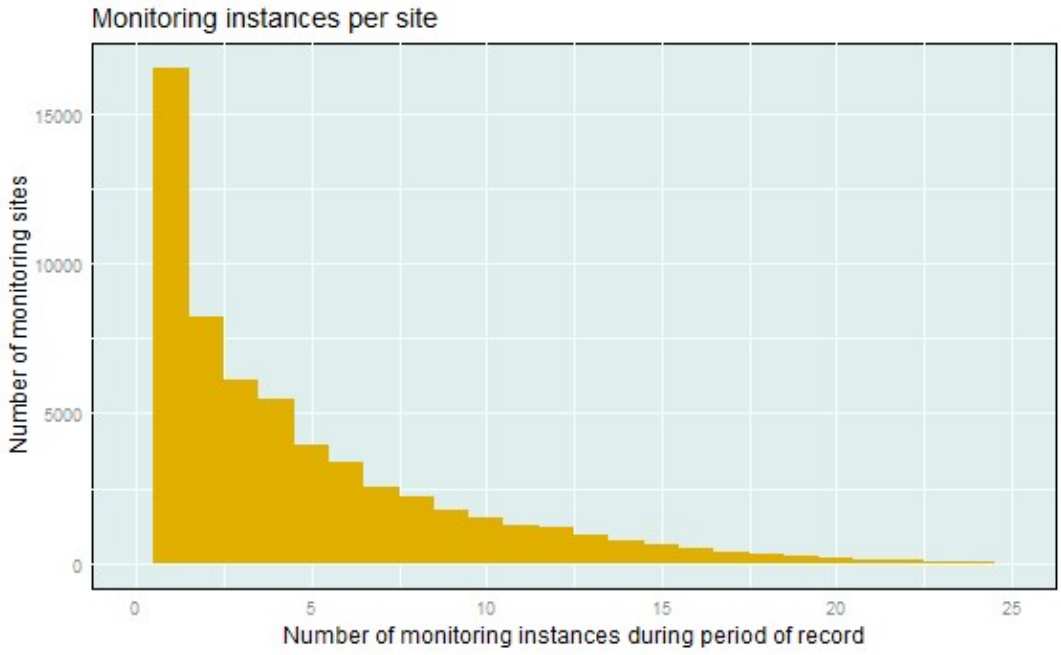

Figure 5 - Monitoring instance frequency of monitoring sites during the period of record.

While each of the 50 states had at least one monitoring site, the degree of reported monitoring varied substantially between states and EPA regions, even when adjusted for land area or population. The number of sites with reported monitoring, broken down by monitored media, is shown for each state in Figure 6 and for each EPA region in Figure 7.

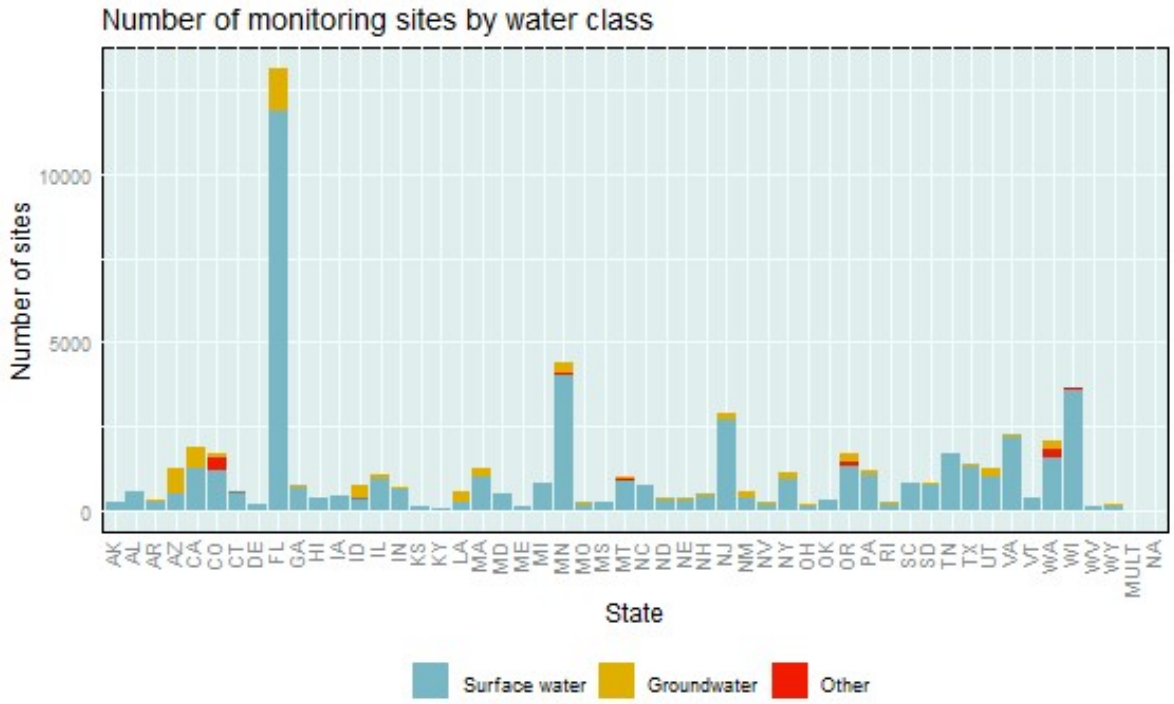

Figure 6 - Number of monitoring sites within each state for which data is reported to the WQP during the period of record, by monitored water media.

The mean number of monitoring sites per state is 1,140. The states with the largest number of sites are Florida (13,120), Minnesota (4,420), and Wisconsin (3,664), while the states with the fewest sites are Kentucky (73), Kansas (108), and Maine (109).

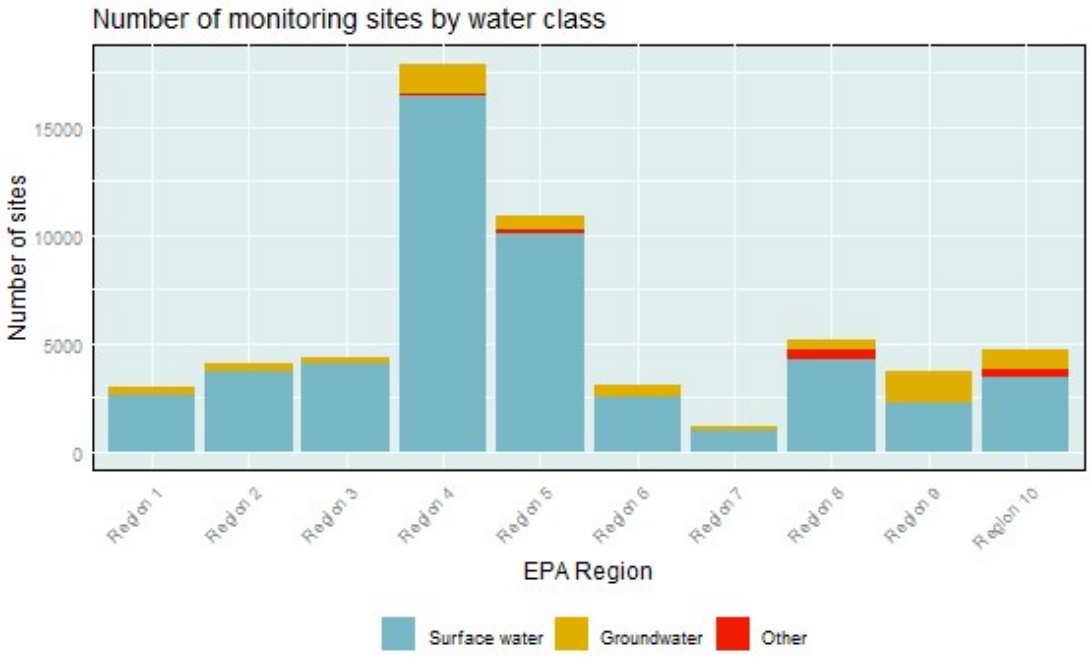

Figure 7 - Number of monitoring sites within each EPA region for which data is reported to the WQP during the period of record, by monitored water media.

EPA Region 4 (serving AL, FL, GA, KY, MS, NC, SC, TN and 6 tribes) had the greatest number of monitoring sites (17,878), while Region 7 (serving IA, MO, KS, NE, and 9 tribes) has the least (1,174). The regional differences are largely an artifact of the levels of monitoring of their constituent states, without a strong signal in the data indicating distinct regional differences.

These differences in the number of monitoring sites also result in different levels of reported subwatershed monitoring. The percent of monitored subwatersheds within a state ranged from 11% (AK) to 90% (NJ) (Figure 8), and from 7% (Region 10) to 45% (Region 3) within an EPA Region (Figure 9).

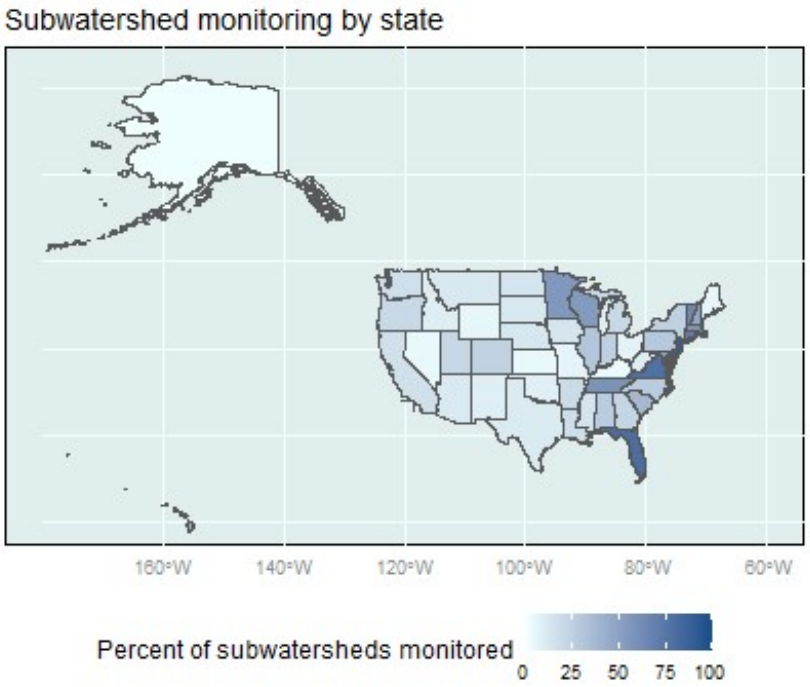

Figure 8 - Percent of subwatersheds classified as monitored by state.

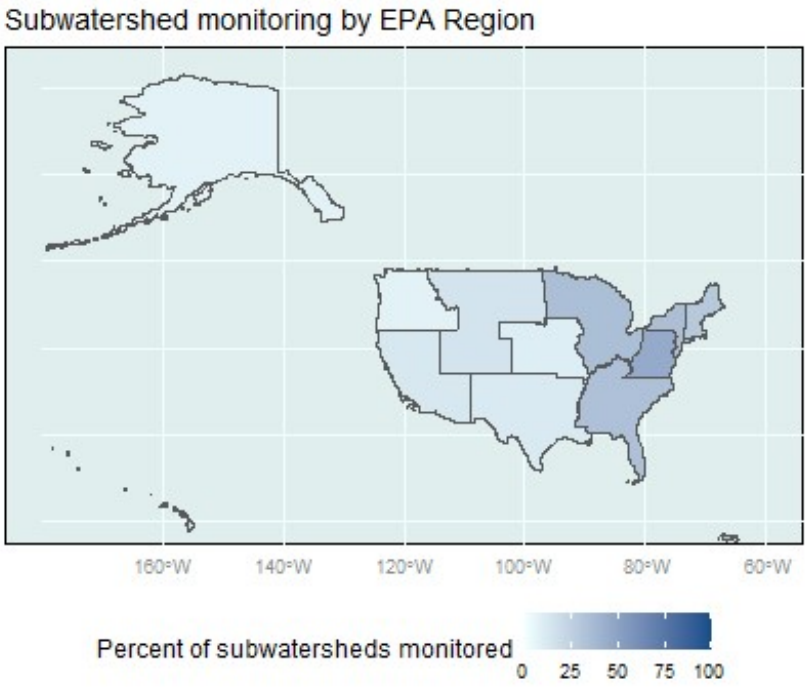

Figure 9 - Percent of subwatersheds classified as monitored by EPA Region.

Because of differences in population distribution and density, as well as in monitoring and reporting, states and regions differed in the proportion of their population served by reported water quality monitoring, ranging from 5% (NV) to 97% (FL) in states (Figure 10) and 31% (Region 7) to 73% (Region 8) in EPA regions (Figure 11). Overall, the proportion of the population served by monitoring was significantly higher than the proportion of subwatersheds served by monitoring, indicating a potential connection between population density and reported water quality monitoring.

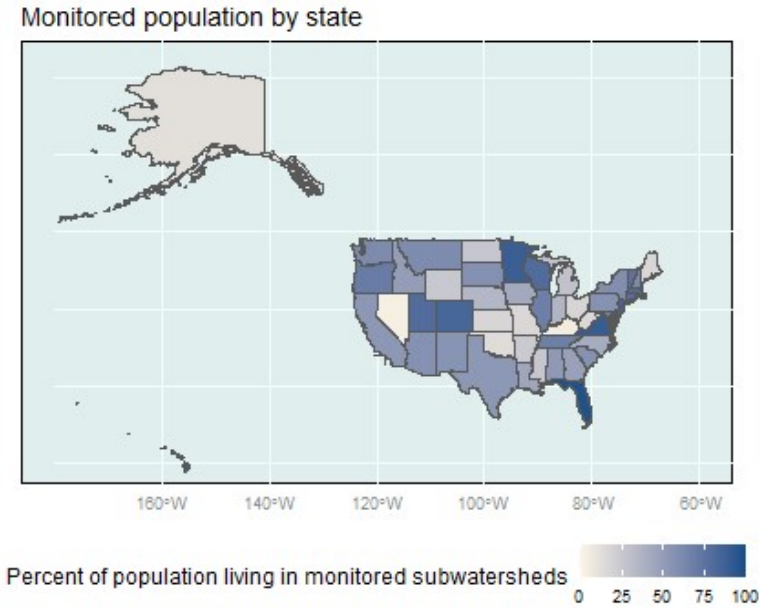

Figure 10 - Percent of population living within subwatersheds served by reported water quality monitoring by state.

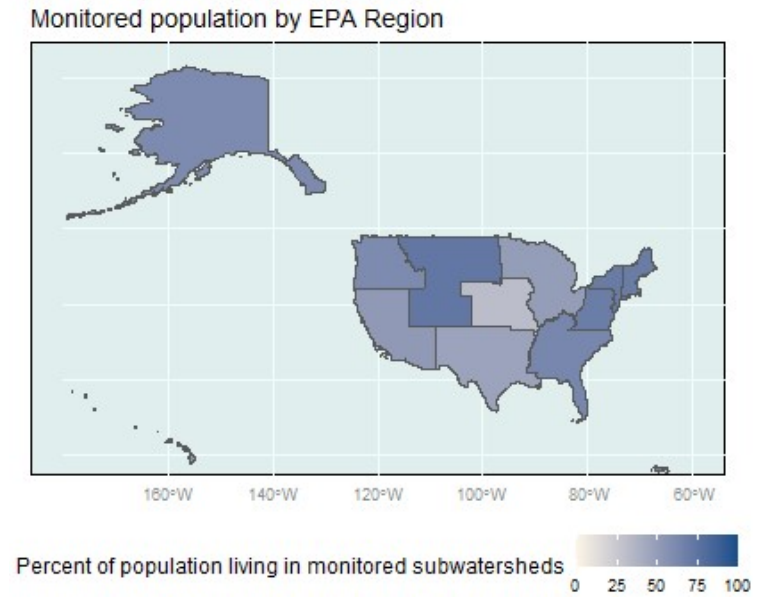

Figure 11 - Percent of population living within subwatersheds served by reported water quality monitoring by EPA Region

PROVIDERS OF DATA

During the period of record, 418 agencies and organizations across 50 states provided data to the Water Quality Portal. The state with the greatest number of providers was Florida, with 66 providers across state, municipal, academic, and private sectors. Kentucky, Missouri, Rhode Island, Vermont, West Virginia, and Wyoming each have a sole provider of water quality monitoring data: their state USGS Water Science Center.

State agencies (35%) and tribal organizations (33%) have the greatest representation among the water quality data providers as a proportion of unique data-providing organizations (Figure 12). However, when considered in terms of the number of monitoring sites contributed, state agencies have a far greater representation: 73% of all monitoring sites are provided by state organizations, with tribal organizations providing only 5% of all monitoring sites (Figure 13). This discrepancy may arise because tribal organizations, while abundant, are monitoring smaller land areas than are state organizations. The high representation of state agencies as providers of water quality data is consistent with the legislative mandates requiring states to report water quality data through the WQX, creating incentives for each monitoring and reporting that are further considered in the policy analysis that follows.

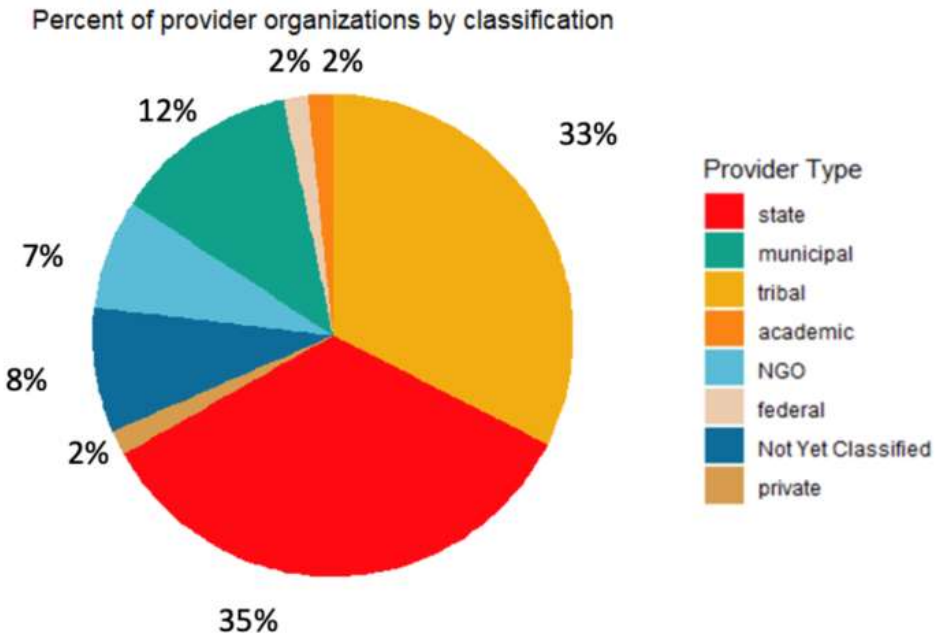

Figure 12 - Proportion of unique data providing organizations belonging to different organization type classifications.

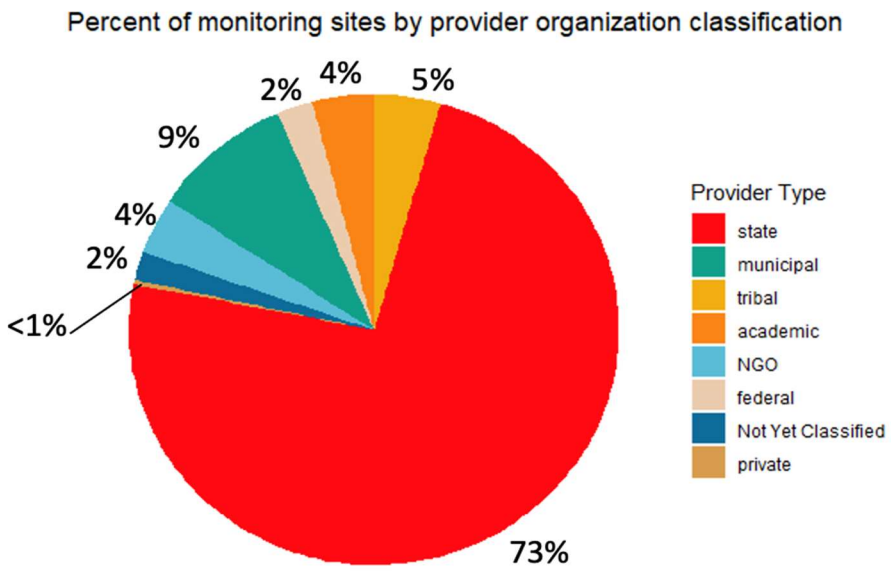

Figure 12 - Proportion of monitoring sites provided by different organization types.

While the precise composition of data providers differs by state and EPA regions, state organizations are the primary data providers across all geographies considered (Figures 14 and 15). Florida stands out for its abundance and diversity of data providers (see Figure 14). This composition may be attributed to state policy surrounding water quality monitoring that requires that the state's water resources are governed collaboratively at each the state and regional level.[11] Regional water resources in Florida are managed through state water management districts which may result in more robust municipal monitoring efforts.

[11] https://floridadep.gov/water-policy/water-policy/content/water-management-districts

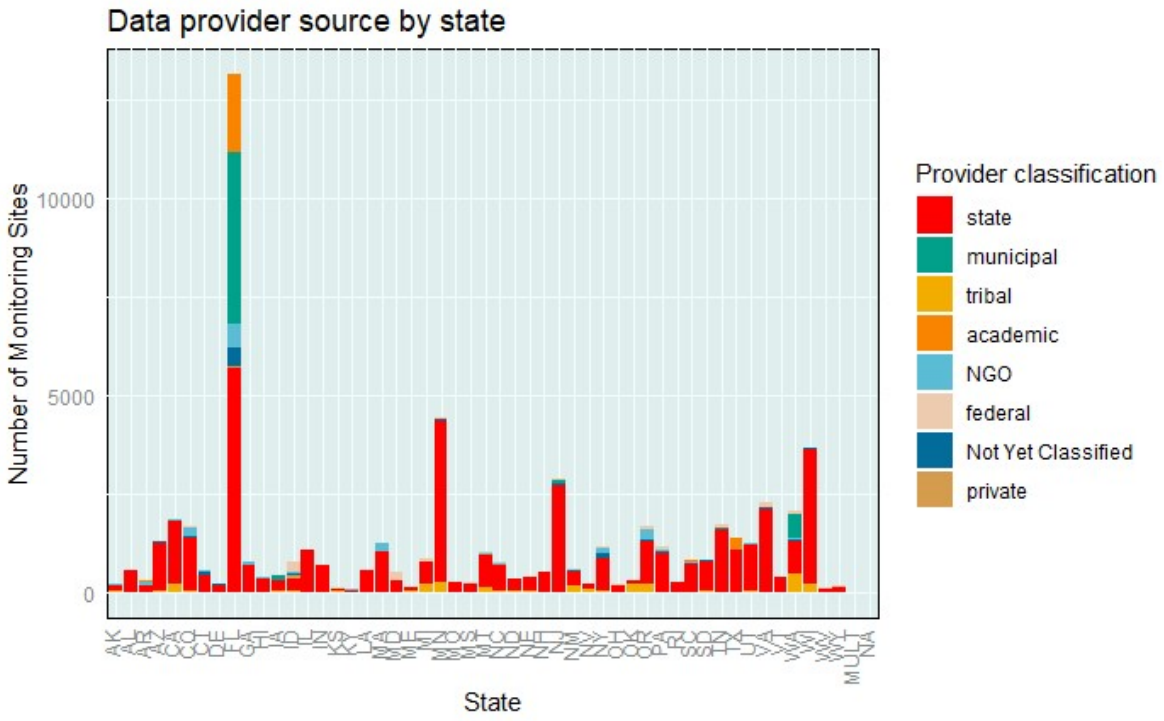

Figure 13 - Number of sites provided by different provider classifications by state.

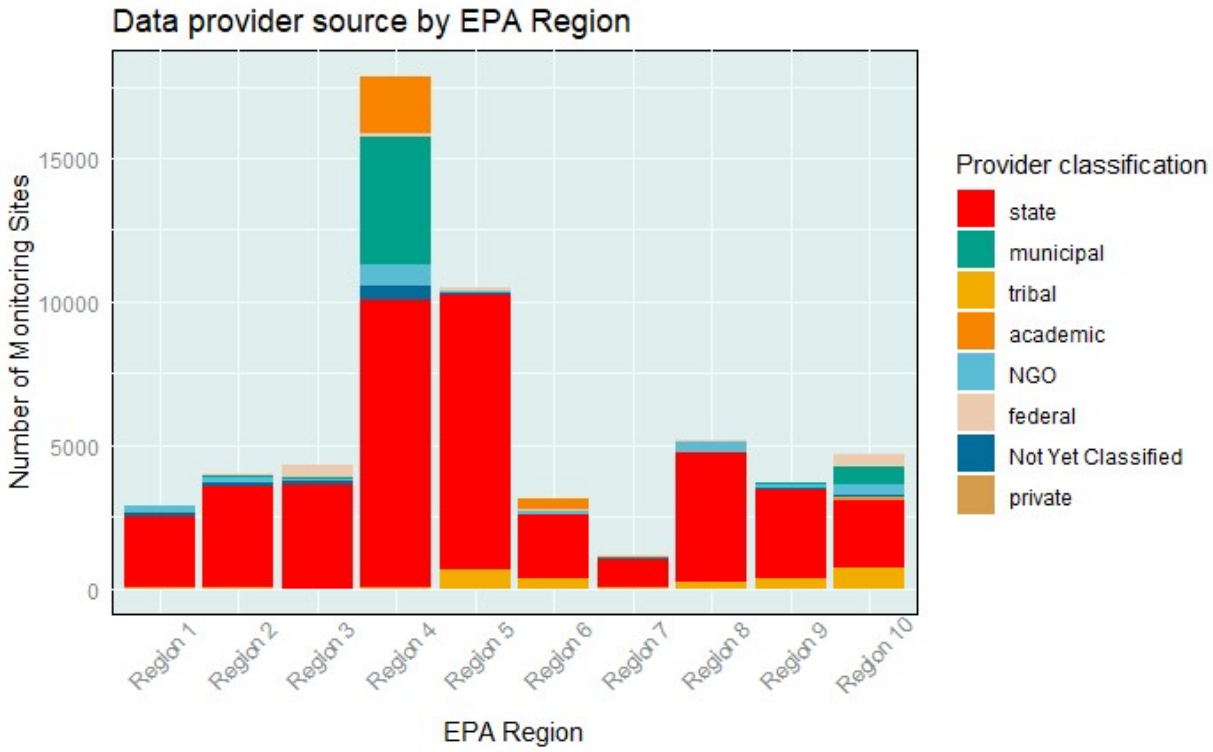

Figure 14 - Number of sites provided by different provider classifications by EPA Region.

MONITORING & INCOME

The median household income of each monitored and unmonitored subwatersheds was examined to determine if household income correlates with water quality monitoring as reported through the Water Quality Portal. The mean of median household income of monitored subwatersheds ($57k) was slightly higher than that of unmonitored subwatersheds ($52k) (Figure 16).

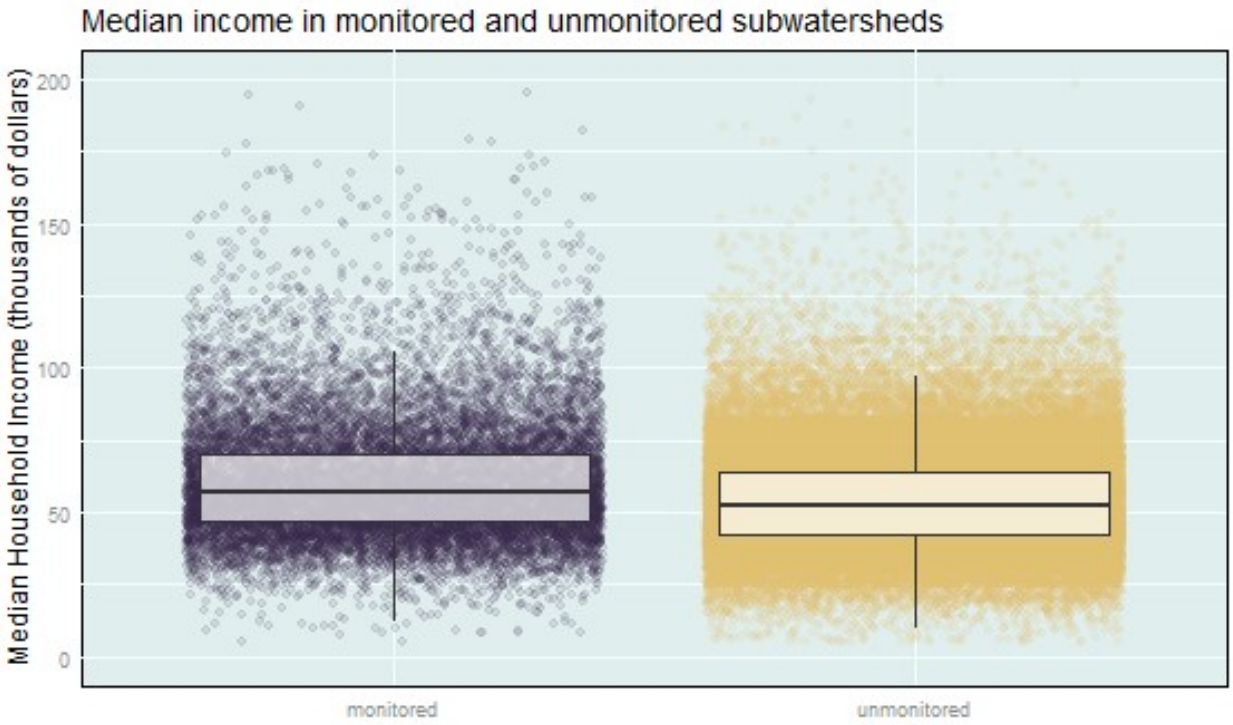

Figure 15 - Boxplot showing the median household income distribution of each monitored and unmonitored subwatersheds. Each point represents the median income of an individual subwatershed.

Higher income communities may be able to support more institutions and organizations with the capacity to collect data, or may be better able to provide the resources needed to report those data through the WQP. However, further inquiry into specific communities and contexts would be needed to understand if there is a causal link between lower incomes and lesser reported monitoring and to thereby devise appropriate policy and practice to address these potential gaps.

Further analysis of economic characteristics of populations served and unserved by water quality monitoring could also consider the poverty prevalence in each monitored and unmonitored subwatersheds to better understand if economically disadvantaged communities are underserved by monitoring data provided through the water quality portal.

MONITORING, RACE, & ETHNICITY

The relationship of monitoring to each race and ethnicity was considered both by analyzing the racial composition of monitored and unmonitored subwatersheds (composition) and by determining what proportion of each racial group lives within monitored and unmonitored subwatersheds (coverage). These metrics allow for potential disproportionate impacts, positive or negative, arising from the spatial distribution of water quality monitoring on different race and ethnic groups.

As a proportion of residents of unmonitored subwatersheds, white and Native American racial groups make up a greater proportion of the population in unmonitored subwatersheds than they do in monitored subwatersheds (74 vs. 66% for white, 1 vs. 0.64% for Native American), which may suggest that these populations are, in places, disproportionately underserved by water quality monitoring collection and/or reporting (Figure 17).

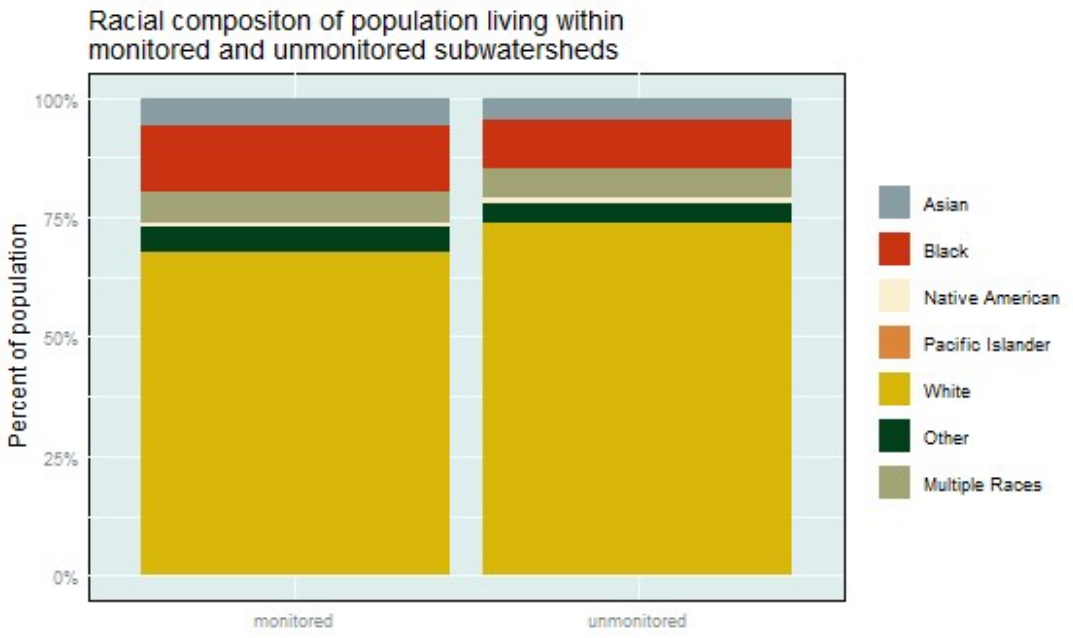

Figure 16 - Racial groups in monitored and unmonitored subwatersheds.

For ethnic groups, non-Hispanic/Latino individuals represent a greater proportion of the population in unmonitored subwatersheds than they do of monitored subwatersheds (84 vs. 80%), while the proportion of Hispanic/Latino individuals is complementarily lower in monitored than unmonitored subwatersheds (Figure 18).

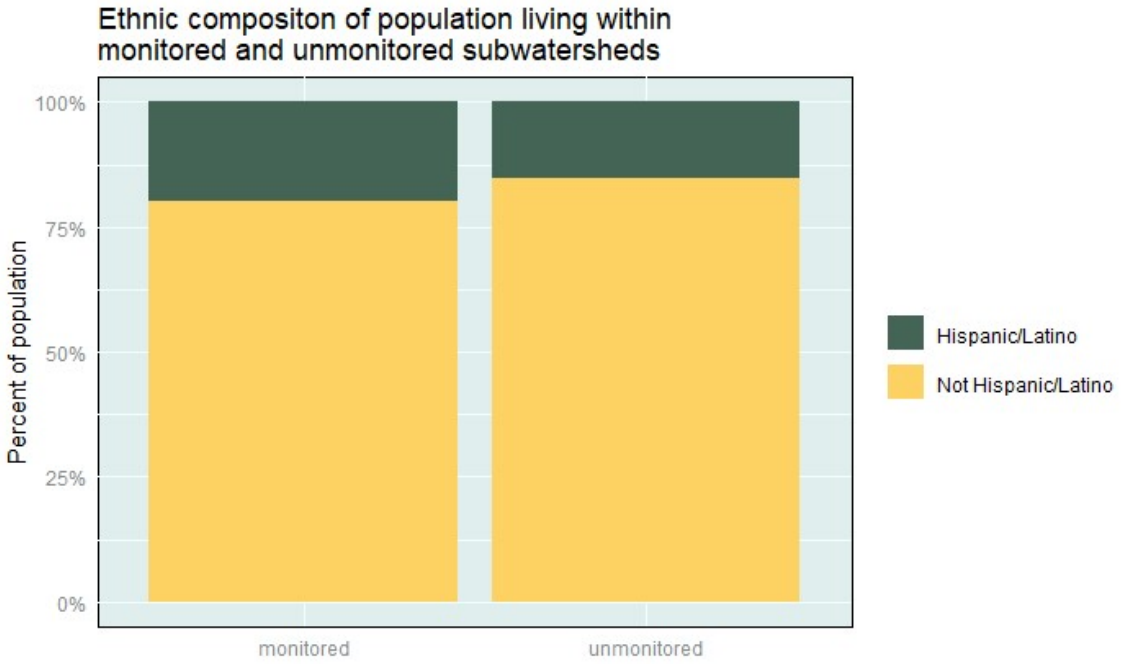

Figure 17 - Ethnic groups in monitored and unmonitored subwatersheds

When considering the distribution of members of each racial and ethnic group that are and are not served by monitoring reported through the WQP, only Native Americans reside in greater numbers in unmonitored than monitored subwatersheds (45% monitored). The racial groups with the highest proportion of the population living in monitored subwatersheds are Blacks, Pacific Islanders, Asians, and those identifying as 'Other' (all 64% monitored), followed by those belonging to multiple races (59% monitored), Whites (55% monitored), and Native Americans (45% monitored, see Figure 19). For ethnic groups, both Hispanic/Latino (63% monitored) and non-Hispanic/Latino (56% monitored) populations live in greater numbers in monitored than unmonitored subwatersheds (Figure 20). This counters our expectations of traditionally understood environmental justice considerations as they relate to racial and ethnic minorities. However, these trends vary by state and region. Furthermore, the presence of monitoring does not negate the possibility of the poorer environmental quality – it simply serves as a mechanism for identifying that this may indeed be the case, as well as providing a possible first step towards remedying the problem. These racial and ethnic trends further suggest the possibility of an urban-rural divide in water quality monitoring and reporting.

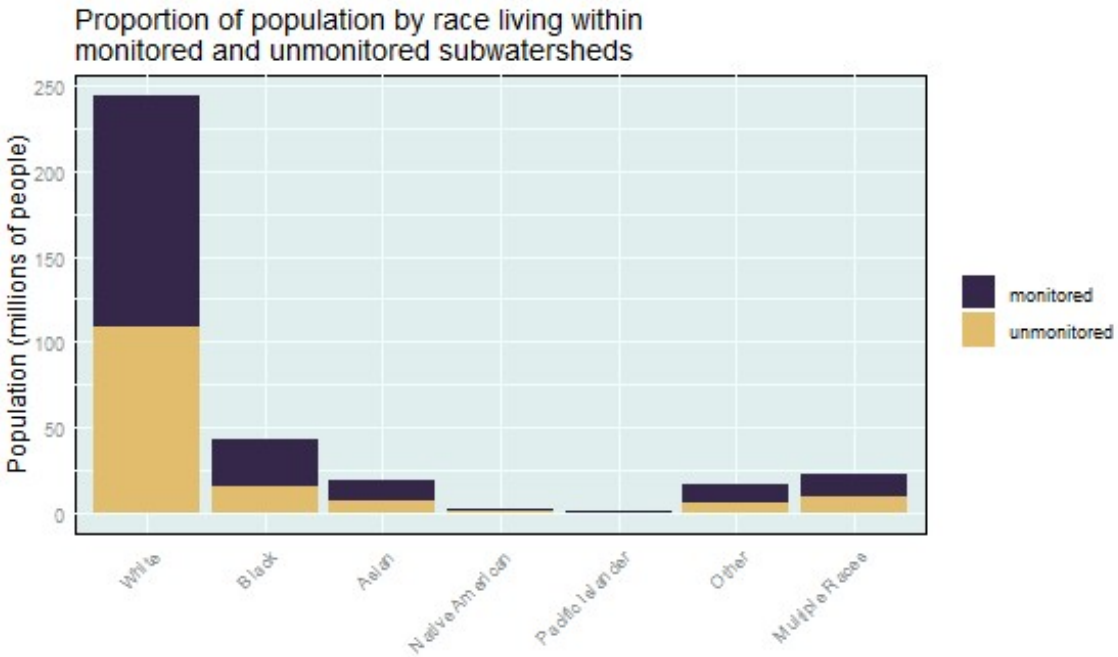

Figure 18 - Population living in monitored and unmonitored subwatersheds by racial group.

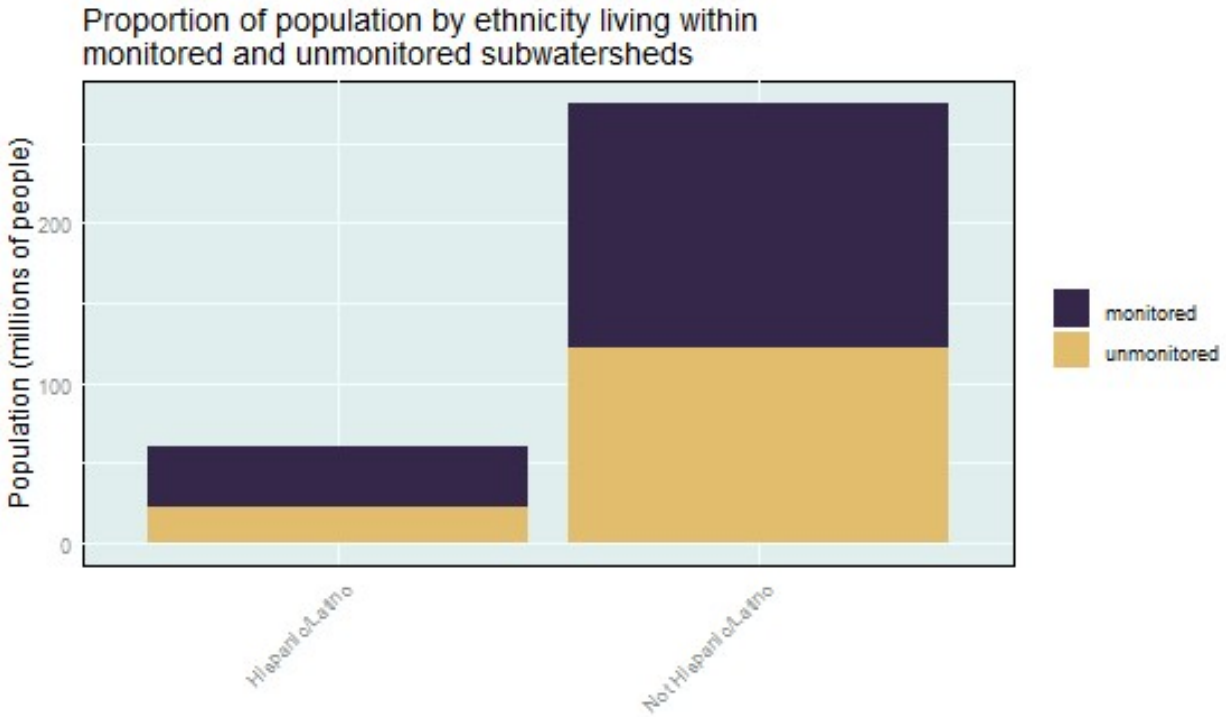

Figure 19 - Population living in monitored and unmonitored subwatersheds by ethnic group.

MONITORING & POPULATION DENSITY

The population density of each monitored and unmonitored subwatersheds was examined to determine if population density correlates with water quality monitoring as reported through the Water Quality Portal. The population density of monitored subwatersheds was, on average, higher than that of unmonitored subwatersheds by more than an order of magnitude (14.3 persons/sq. km in monitored vs. 1.2 in unmonitored, see Figure 21).

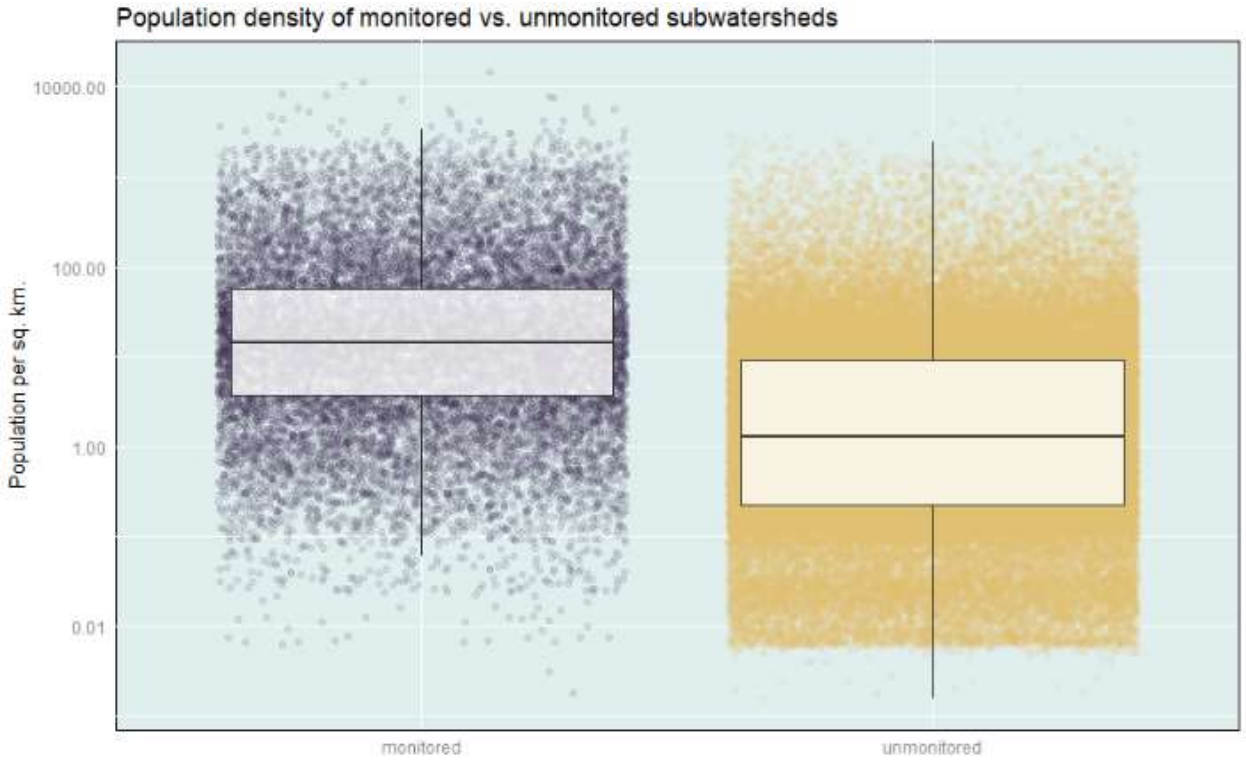

Figure 20 - Boxplot showing the population density distribution of each monitored and unmonitored subwatersheds. Each point represents the population density of an individual subwatershed. Note the logarithmic y-axis.

A higher incidence of monitoring in regions with higher population makes sense, as there are both more people who would be benefitted by water quality monitoring (higher return on investment) and a greater chance that there exist organizations or individuals with the capacity to perform water quality monitoring in these areas. Nonetheless, this pattern also suggests that there may exist an urban-rural divide in water quality monitoring data collection and reporting. This possibility is further examined in Figure 22, where the spatial distribution of regions of each high and low monitoring and high and low population, and the coincidence of these, is shown.

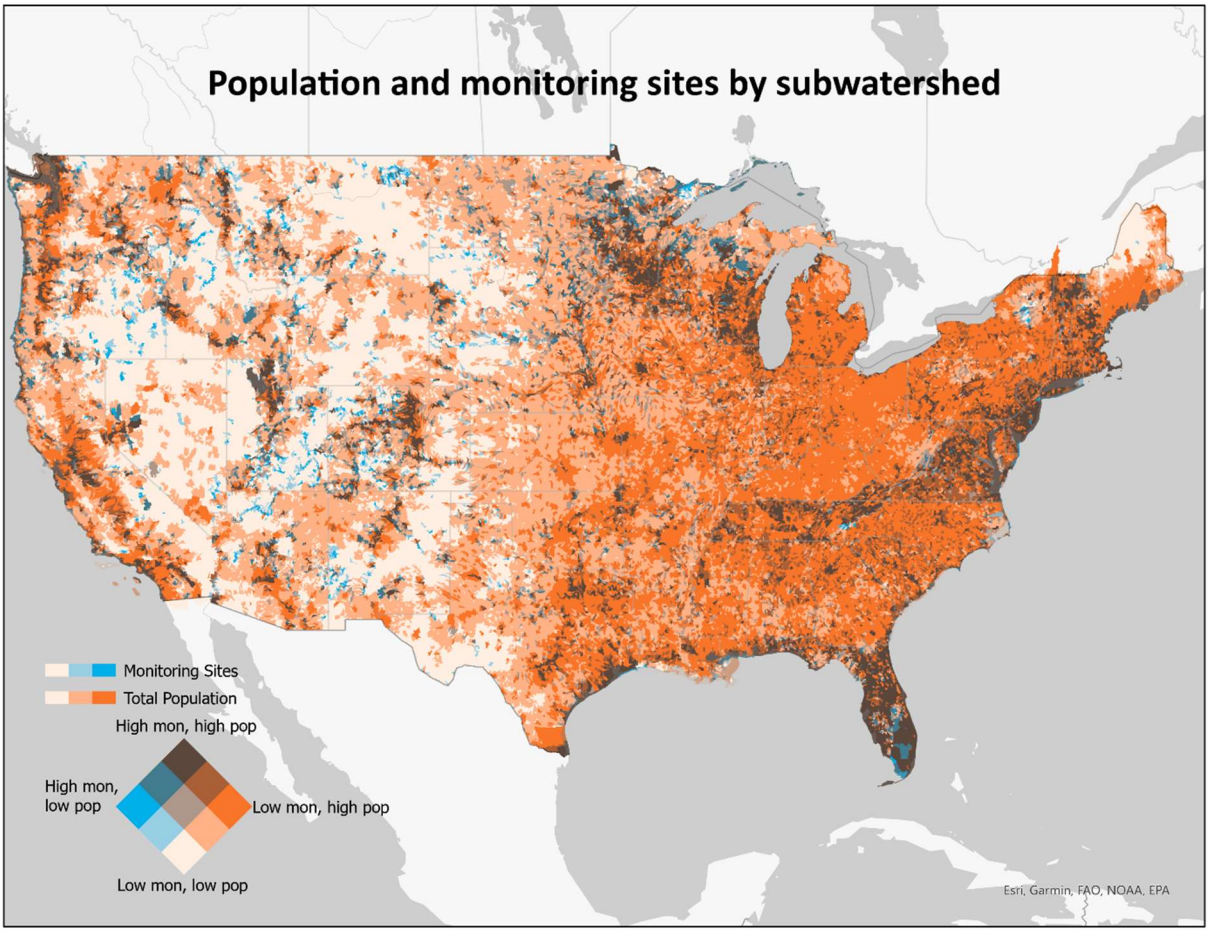

Figure 21 - Bivariate choropleth map showing the spatial coincidence of water quality monitoring sites and human population. A version of this map is included in the project dashboard.

It is unsurprising that the eastern United States has more subwatersheds with high population and low monitoring than does the western United States, where population density is lower. Even so, further investigation into regions falling on any of the corners of the bivariate key could be illustrative, revealing how and why human populations are or are not served by water quality monitoring.

Population density can serve as a proxy indicator of the urban-rural status of each subwatershed. A more precise investigation into the urban-rural divide in water quality monitoring suggested here could be conducted in future work by integrating the incorporated municipal places dataset available through the census bureau.

Since the WQP portal is a fairly comprehensive record of water quality monitoring activities within the United States, the observed gaps and discrepancies may indicate opportunities where increased water quality data collection would be beneficial. However, since the WQP is not a *wholly* comprehensive record of all water quality monitoring activity, they may also indicate regions where, despite the presence of data collection, increased reporting and integration of water quality data into the WQP may be required to achieve the full benefits of water quality monitoring data.

Understanding the full implications and severity of the gaps and disparities reported above will require a more precise specification of the connection between the level and nature of water quality monitoring and real-world outcomes for the health and wellbeing of human and aquatic life and communities. It is our hope that this analysis and the corresponding dashboard can begin to raise, and aid in answering, these questions for individuals involved in decision making about water quality monitoring and reporting, and the landscape of policy and practice. We next offer some recommendations for policy and practice to increase the coverage and comprehension of the data available through the water quality portal.

POLICY ANALYSIS AND PROPOSALS

Non-existent or insufficient water quality monitoring as reported by the data contained in the WQP could arise from two primary causes: it may indicate a lack of monitoring, or it may indicate a lack of reporting. In either case, the end result of data unavailability is the same, but the solutions offered by both policy and practice will differ depending on the reason for a lack of data. To minimize the spatial gaps in water quality monitoring, there must be action on both the state and federal level to ensure frequency in both monitoring and reporting along with sufficient resources to ensure that states have the capacity to help providers report their data.

This analysis examines the existing legislation salient to water quality monitoring and reporting practices, explores best practices in the dissemination of water quality monitoring data, and provides solutions that aim to eliminate gaps in the provision of water quality monitoring data.

POLICY LANDSCAPE

Federal mandates administered by the EPA currently require that all states adhere to water quality monitoring and reporting practices. Under Sections 303(b) and 303(d) of the Clean Water Act, states must assess waterbodies and list surface water bodies that are polluted.[12] This Clean Water Act legislation requires states to prepare an Integrated Water Quality Monitoring and Assessment Report every two years. The objective of this report is to assess the quality of statewide waters (lakes, wetlands, and streams), identify impaired waters, and report on the number of pollutants present by developing Total Maximum Daily Loads (TMDLs). Water quality monitoring data that is reported to the EPA in conformance with these requirements is shared through the Water Quality Exchange (WQX), making it publicly available on the WQP.

While federal mandates help guide water quality monitoring and reporting on the state level, states have latitude in the monitoring and reporting practices implemented through state-level water quality monitoring strategies. These strategies generally summarize the state's water quality monitoring activities and goals and describe how the goals outlined will be achieved, including partner organizations whose data may be incorporated into state monitoring reports. While some states publish an annual water quality monitoring strategy, others prepare these strategies every five to ten years. The EPA requires that state water quality monitoring strategy timelines not exceed ten years.[13]

Since federal mandates require that states share their water quality monitoring data with the EPA to fulfill Clean Water Act guidelines, state water quality monitoring strategies also discuss the procedures and databases in place that enable states to submit data to the EPA. Some states, such as California, have comprehensive water quality reporting measures, as with the California Environmental Data Exchange Network (CEDEN). CEDEN is the California State Water Board's data system for surface water quality monitoring in the state.[14] This data system helps ensure frequent water quality data reporting in California because it provides state agencies and tribal organizations and NGOs with a centralized platform to report data. In addition, CEDEN shares statewide data with the WQX so that it is publicly available in

[12] https://www.waterboards.ca.gov/sandiego/water_issues/programs/303d_list/).
[13] https://archive.epa.gov/water/archive/web/html/elements.html#N_8_
[14] http://ceden.org/about_us.shtml

the WQP, integrating data from a variety of provider organization types into the data provided through the portal.[15] California's water quality data reporting practices demonstrate how states can develop systems to ensure data from a variety of providers is integrated into the WQP. On the other hand, states such as New Hampshire face barriers in incorporating their data into the WQX. While the New Hampshire Department of Environmental Services (NHDES) aims to submit water quality monitoring data to the WQX within two years of collection, the NHDES faces challenges with integration because their data collection form is not compatible with the WQX.[16] This obstacle coupled with a limited data management staff have made it more difficult for New Hampshire to regularly share any data beyond those mandated for reporting. Examining water quality reporting practices such as those in California and New Hampshire demonstrates how data-sharing programs can facilitate reporting and how a lack of capacity can impede reporting.

ADDRESSING A LACK OF MONITORING

Gaps in the data reported through the Water Quality Portal may indicate a true absence of water quality monitoring. Changes to policy and practice can increase the feasibility of, and incentives for, monitoring.

The existing Clean Water Act regulations pertaining to water quality require states to perform ambient water quality monitoring on an annual basis, though states are responsible for the precise interpretation and implementation of that monitoring. Not all water bodies are monitored for this: rather, a representative subset of state water bodies are surveyed to provide insights on all the bodies of water in the state. States therefore differ in their level and patterns of water quality monitoring and reporting. Our analysis shows that there was not a consistent trend between land area or population and level of monitoring in our analyses, and only a weak correlation between state revenue and the percent of state population that lives within monitored subwatersheds. A further correlational analysis of the proportion of subwatersheds or population served by water quality monitoring that considers aspects such as flow line length, waterbody area, state environmental budget, or water quality impairment, among other possible factors, could be insightful. The rationale behind the level of reported monitoring may also require consideration of state or region-specific factors such as tourism associated with water resources or historical bad press related to water quality.

While state organizations are strongly represented in water quality monitoring due to federal mandates, water quality monitoring across other organizations types (ex. municipal, tribal, NGO, academic) are not as prevalent, partially due to the absence of perceived incentives or value, legislative or otherwise. While legislative mandates for monitoring and reporting begin to provide important insights into water quality, this monitoring often only occurs on a representative sample of state waterways. More comprehensive coverage may enable communities to better understand the conditions of their waterways and determine whether their water is safe for consumption or recreation, and non-state organizations can meaningfully contribute to collecting data beyond what is currently collected or required by legislation.

Building out existing grant programs such as the CWA Section 106 (Water Pollution Control Grants) can help provide organizations with greater institutional capacity to monitor and report on the quality of waterways. Water quality activities eligible for Section 106 funding include but are not limited to

[15] https://www.sfei.org/news/ceden-update#sthash.LBPfNyUm.dpbs
[16] https://www.des.nh.gov/sites/g/files/ehbemt341/files/documents/r-wd-16-02.pdf

performing water quality monitoring and assessments, developing a monitoring strategy, and hiring staff to identify and prioritize water quality issues.[17] These grants typically range from $40,000-$200,000, and first-time applicants are limited to grants for $40,000.[18] Increasing funding for Section 106 could ensure more organizations receive support for water quality monitoring activities. Another step in bolstering this grant program is for states to provide more guidance on Section 106 funding to water quality monitoring organizations. The EPA currently outlines the application process and steps for receiving grant funding; however, state environmental agencies can further that guidance by incorporating Section 106 funding guidelines into their water quality monitoring strategies or on their websites to enable data providers to have better access to those resources.

States can also incentivize data collection by other organizations by ensuring that these data will be meaningfully considered and incorporated into state management and funding plans. Would-be data providers may lack the resources to address any water quality issues that monitoring would identify, and may therefore not perceive the monitoring to be a valuable activity. By partnering with local agencies and NGOs, states can bolster the level of monitoring through dispersed efforts while ensuring the benefits of water quality monitoring are maximized. Developing plans and programs to provide uniform Quality Assurance/Quality Control (QA/QC) for these efforts will help ensure their credibility.

ADDRESSING A LACK OF REPORTING

Gaps in the data provided through the Water Quality Portal may also indicate a failure to report or integrate data that are already being collected into the WQP. Non-state organizations may collect water quality data for a variety of monitoring, educational, and research purposes, but these data may not be reported publicly or integrated into the WQP.

For water quality monitoring data to be publicly available through the WQP, data providers must ensure that their data is compatible with the WQX-format. The WQX data format provides standardized best practices, domains, and formats to ensure the inter-operability of data generated by a diversity of providers. This standardization is key to the broad utility and comprehensive nature of the Water Quality Portal. However, the same standardization that makes the WQP a valuable resource can be an impediment to data reporting, as it requires compatibility between in-house data management methods and the WQX. This can be accomplished either by adopting WQX methods for in-house data management or by reformatting the in-house formats for upload.[19] These options can be complicated, time consuming, and expensive. While there are processes to ensure data is integrated into the WQX, data providers may lack the resources, capacity, or technical capabilities and personnel to regularly reformat the data.

One option for maintaining data in a format compatible with the WQX is the Ambient Water Quality Management System (AWQMS), a proprietary water data management software developed by Gold Systems, Inc. [20] Because Gold Systems also developed the WQX format, data managed through AWQMS are easily integrated into the WQP with one click. However, the software's annual fees can be a financial barrier for organizations ($3,320 annually for an individual tribal or volunteer organization). [21] Grant

[17] https://www.epa.gov/tribal-pacific-sw/r9tribalcwa
[18] Ibid.
[19] https://www.epa.gov/waterdata/water-quality-data-upload-wqx
[20] https://www.awqms.com
[21] https://www.awqms.com

funding through CWA Section 106, which requires recipients to upload their data to the WQX, could be used in part to cover the licensing costs associated with AQWMS. AQWMS licensing is also available for consortiums of data providers, with decreased costs for larger collaboratives, which can ease the financial burden. Organizations can also pursue collaboration to share the technical and personnel resources needed to manage data in a manner consistent with regular WQX integration, as centralizing efforts to upload data can help overcome barriers posed by a lack of technical expertise and capacity by allowing multiple local or regional organizations to share data management resources and costs associated with integrating data into the WQX/WQP. Reducing the barriers of cost, time, and expertise associated with reformatting data from in-house data formats to be compatible with the WQX could improve data upload and integration rates.

Building partnerships and collaborative processes across organizations to address water quality data goals can further bolster monitoring and reporting initiatives. A noteworthy example of effective water data collaboration is California's Freshwater Harmful Algal Blooms (FHAB) Monitoring Program. The FHAB program goals are to provide assessment, response, and guidance on statewide harmful algal blooms.[22] This program relies on collaboration and consultation with state and federal agencies and state tribal organizations to implement monitoring and reporting goals while ensuring that data is submitted to the WQX. The FHAB program can provide a model of effective collaboration in the collection and reporting of data for improved water quality monitoring outcomes.

State and federal policy efforts to promote open water data initiatives can also help improve water quality data reporting. Currently, California, New Mexico, and Oregon have enacted legislation that will dedicate capacity to developing and improving statewide water databases. Oregon's new legislation demonstrates a commitment to improving data integration through establishing investments that will allow the Oregon Department of Environmental Quality to create a water data platform that can integrate multiple databases and enable the public the retrieve water quality data through one interface.[23] While water data legislation can take on multiple shapes and forms, one common theme seen across these policies is the commitment to making data more accessible to the public by eliminating barriers in reporting and integration. The further step of ensuring all publicly reported data are integrated in to the WQX and WQP will allow for greater ease of discovery and use of water quality data across the United States. Promotion of these open data initiatives can play a key role in increasing the level of data collection and reporting, increasing the comprehensiveness of data provided through the WQP.

[22] https://www.waterboards.ca.gov/water_issues/programs/swamp/freshwater_cyanobacteria.html
[23] https://www.oregonlegislature.gov/courtney/Documents/2021-Water-Package-Release.pdf

DISCUSSION + AVENUES FOR FUTURE INQUIRY

This project sought to provide a preliminary investigation into the extent of the water quality monitoring represented by, and reported through, the water quality portal. It also sought to understand who is responsible for providing these data as well as to identify what populations or communities may be underserved by these monitoring efforts.

As a preliminary investigation, several assumptions and simplifications were made for the sake of analysis. These include:

- The spatial scale for which water quality monitoring data are valid was assumed to be the subwatershed (HUC12) within which they are located as well as any subwatershed immediately downstream of that subwatershed. The actual spatial extent and downstream distance for which water quality monitoring data provide valid and useful information about a body of water may differ and may depend on the parameters in question and the specific context of a given locale. For example, perhaps some monitoring data could pertinently indicate the conditions of an entire subbasin (HUC8) or may impact water quality three subbasins downstream, while others may be too localized to even extend to the entire subbasin within which they are located.

- The temporal scale for which water quality monitoring data were considered to be valid was 2 calendar years after their collection. The actual period over which water quality monitoring data provide valid and useful information about a body of water may be longer or shorter, and may depend on the parameters in question and the specific context of a given locale, season, or year. Because of this, and given that the vast majority of the data collected for this project occurred during the COVID-19 pandemic, the period of record considered in this analysis may not be representative. It is possible that water quality monitoring and reporting occurred less frequently during this time due to the disruptions caused by the pandemic.

- All monitoring sites were considered to provide the same extent/quality of monitoring, regardless of the frequency with which they monitored, the time since monitoring within the selected period of record, or the number or nature of water quality parameters measured. The inclusion in the monitoring site dataset presented on the dashboard of the number of the monitoring instances, the monitoring frequency classification, and tallies of the measured parameters by regulatory grouping at each site attempts to allow dashboard users to bring their own perspectives on the importance of each of these into their interpretation of the data and subsequent decision making.

- Any monitoring site within a subwatershed or the subwatershed from which it drains was considered to render that subwatershed ‘monitored’, without accounting for whether the site was located along the actual reach of stream around which the subwatershed is defined. Thus, monitoring occurring on a hydrologically disjoint pond would nonetheless result in a subwatershed being classified as monitored, even though it does not directly pertain to the quality of the main flowline. The same is true for groundwater monitoring sites.

- In many of the analyses presented above and in the dashboard, there is a binary consideration of monitoring at subwatershed scale: that is, a subwatershed would be classified as "monitored" whether it contained one monitoring site with one instance of temperature data collection occurring 2 years ago or if it contained 50 monitoring sites recording data every month for a wide variety of parameters over that same period. As noted above, details about the nature and extent of monitoring occurring at each sites are presented with the monitoring points dataset included in the dashboard to allow users to delve into the details in their region of interest.

- The populations served (or unserved) by monitoring were considered to be those living within a given subwatershed, as interpolated from census geographies coincident with those subwatersheds. In addition to the assumption of homogenous distribution of individuals and households within census geographies, the analysis also assumed that individuals are only affected by the water quality monitoring of the subwatershed in which they live. This is a simplification, as individuals may work in, travel to, or receive drinking water from subwatersheds other than those in which they reside, and accordingly be affected by the presence or absence of adequate monitoring in these other subwatersheds.

This project does not, and was never intended to, offer a definitive nor conclusive statement about what populations are or are not adequately served by water quality monitoring. Rather, it aims to provide preliminary insights into the spatial patterns of water quality monitoring as indicated by data availability through the Water Quality Portal, and to examine what human populations are (not) served by these monitoring activities. The analyses and discussion presented here aim to highlight potential trends and patterns and to identify opportunities for future work, discussed below.

FUTURE INQUIRY

Potential refinements and extensions of the analysis at hand are presented below. The code and resources contained in the GitHub repository linked above can serve as a starting point for continuing the work begun in the present project.

- Further work could investigate the appropriateness of each of the assumptions presented above, considering the sensitivity of the results presented here to the spatial, temporal, and parametric conceptions of monitoring validity employed in the present analysis.

- Other population or community characteristics could be examined in relation to monitoring to determine if certain populations are being underserved by water quality monitoring and to develop correspondingly targeted recommendations for policy and praxis. These could include sex, language, home tenure and ownership, and municipal incorporation.

- Work focused specifically on equity and justice could adapt our methodology to analyze the relationship between monitoring and income by focusing on low-income households rather than on median household income.

- ➢ The dashboard could be further developed to allow for more precise filtering of the site characteristics used to determine whether a given subwatershed is considered to be monitored, allowing users to refine and customize the analysis for their particular questions.

- ➢ Further investigation into the role of, and benefit to, tribes in providing water quality monitoring data could provide insight into strategies for increasing the monitoring coverage of Native American individuals. An explicit consideration of tribal boundaries could prove helpful in this.

- ➢ An investigation into the relationship between impaired waterways and water quality violations and monitoring could provide insight on how a prevalence or lack of water quality monitoring has implications for the health and well-being of human and aquatic life.

- ➢ Access to waterways for monitoring may be limited by private ownership of surrounding lands and by the differing legal rights-of-way afforded to waterways by states. An investigation of the proportion of private vs. public property in subwatersheds could provide insight into the extent to which private property ownership restricts water quality monitoring.

- ➢ Explicit consideration of state and regional budgets, policies, and programs could provide valuable insight into what policies and practices have proven most effective in promoting robust water quality monitoring and reporting.

- ➢ An investigation into other water quality data repositories maintained by state, local, and other organizations could provide insight into how comprehensive the monitoring data available through the WQP are and identify where monitoring is occurring that is not reported to the WQX nor available through the WQP. Interviews and stakeholder engagement with producers of water quality monitoring data that are not represented in the WQP could then highlight potential barriers preventing the upload of data to the WQX, in turn pointing towards practical solutions that could be implemented to improve the efficacy and inclusivity of the Water Quality Portal.

The existence and availability of monitoring data are important first steps in ensuring water quality. However, data availability alone is not sufficient: the presence of monitoring is no guarantee of the quality of the measured environment. Water quality monitoring very well may indicate poor environmental quality without directly translating into action or improvement. Nonetheless, where monitoring does not occur, even the possibility for identification of and action upon causes for concern is absent. Even where monitoring does occur, if its findings are not publicized or accessible, its usefulness for identifying and responding to water quality concerns will be limited. Thus an analysis of the extent and distribution of monitoring data is a valuable step towards ensuring water quality, but not in itself sufficient. Rather, it ought to inform further work to increase monitoring data collection, improve data accessibility and transparency, and appropriately consider and address water quality issues indicated by these data. By doing so, water quality monitoring can be a crucial tool in protecting and promoting the health and wellbeing of human and non-human communities alike.

ACKNOWLEDGEMENTS

Special thanks to Ashley Ward, Martin Doyle, Lauren Patterson, and Kyle Onda at the Internet of Water for providing the idea for this investigation and for thoughtful feedback, guidance, and encouragement along the way.

APPENDIX: ANALYSIS METHODS

All data access and wrangling was accomplished through R, with full code available at the following link:

https://github.com/sab159/MP_SpatialPatternsWQPData.

There are three scripts used for creating and updating the data that support the dashboard: A script to (where necessary, install and) load the packages used in the workflow and to configure the global environment, to be run before either of the other scripts; a script to pull and compile the initial data that will support the dashboard, to be run once to set up the dashboard inputs; and a script to update the dashboard data, to be run occasionally to update the data presented in the dashboard. The update script also includes some data analysis and enhancement and should be run every time the initial script is run. An R Markdown file is also included for analyzing and visualizing the data and can be used to reassess the analyses presented in the findings presented in the main body of the text after updates to the data. A Python script is included that updates the hosted feature layers powering the dashboard with the files created locally by the other scripts. The following sections discuss the methods used for accessing, compiling, updating, and analyzing the data.

WQP SITES

Metadata

Metadata and data for all monitoring sites represented in the WQP were accessed using the 'dataRetrieval' package for R. There were, in total, 2.6 million sites at the time of access in March 2022. Site metadata included information about the organization collecting/reporting the data, the unique ID of the monitoring location, categorical information about the site's location (state, county, HUC8, aquifer), and spatial location of the site (longitude, latitude, and the coordinate reference system). There were 11 distinct coordinate reference systems, as well as three "catch-all" categories ("OTHER", "Unknown", and "UNKWN"). Of the 11 coordinate reference systems, 8 were obscure coordinate reference systems that represented fewer than 35 sites each. The remaining three identified coordinate reference systems were NAD27, NAD83, and WGS84. The spatial locations for all sites were reprojected into NAD83, as this was the native coordinate reference system of the greatest number of sites. A significant number of sites (95000+) listed the coordinate reference system for the provided latitude and longitude as 'Unknown/UNKWN'. For these sites, the coordinate reference system was set to NAD83 – this assumption may introduce some spatial error. The lack of state-plane specification for use with the NAD83 coordinate reference system within the original data also introduces some spatial error and uncertainty.

Site Characteristics

Water quality monitoring coverage is determined by both the spatial and temporal distribution of reporting. For this analysis, sites and monitoring records were excluded if their most recent data were collected more than two years prior to the first day of the current year. This functionally assumes that data cannot be considered as providing monitoring coverage after two years, though it does not intend to thereby positively imply that data collected within the past two years *do* provide adequate monitoring coverage: The adequacy of this assumption will vary based on specific water quality parameters, as well as the social and environmental contexts for within which monitoring data are being collected and used. While primarily serving water quality data, some monitoring records provided by the WQP pertain to measurements of air, soil, and other

media: Only those records pertaining to water were retained. Records whose results had been flagged as "Rejected" were also omitted. At the time of development, this resulted in the retention of 292842 monitoring instances at 58148 unique sites. Monitoring instances were aggregated by site, and retained sites were then analyzed for the type and frequency of their monitoring, the categorization of their data provider, and the water quality parameters on which they report.

Sites were classified as either discrete or continuous based on the frequency of their monitoring activities. Sites were considered continuous if they have been monitoring for at least three months (defined by a date of first monitoring during the period of interest occurring more than 95 days ago), if they were currently reporting monitoring data (defined as reported monitoring within the past 35 days), and if they were reporting data at least once a month (max difference between monitoring instances of 35 days, to accommodate potential delays that may spread monthly sampling over a slightly longer period). Sites that did not meet these criteria were considered discrete.

Data providers were manually classified, identifying each unique monitoring-conducting organization as federal, tribal, state, municipal, NGO, academic, or private. Some organizations, such as commissions formed for the administration of inter-state compacts, did not fit neatly into these categories – in such cases, the highest-level classification for any contributing organization was assigned (in the case of compact commissions, federal). As new providers submit data to the portal, manual update of provider classifications may be needed on occasion to keep the dashboard data complete and current. Between updates, providers for whom no classification exists are labelled as "Not yet classified". These provider classifications were appended to the metadata summary records for each site.

Water quality parameters were manually classified according to their pertinence to Clean Water Act (CWA) and Safe Drinking Water Act (SDWA) standards. As with the provider classifications, the manual classification of parameters means that occasional updates will be required to keep the dashboard complete and current. Between updates, unclassified parameters will be aggregated in the "other" category. CWA parameter classifications, based on the EPA's recommended criteria for implementation by states, were subdivided into those concerned with human health[24], with the health of aquatic life[25], and with organoleptic properties[26]. SDWA parameter classifications were similarly subdivided according to primary drinking water regulations[27] and secondary drinking water standards[28]. Categories were also included for basic water quality parameters and for all other parameters represented in the monitoring data but not corresponding to any other category. A single parameter could be included in multiple categories. The number of unique parameters measured within each category at each monitoring site during the period of interest (past 2 years) was then calculated.

Once these calculations and categorizations had been made, monitoring instances were aggregated into a single record for each monitoring site, containing the classifications of the

[24] https://www.epa.gov/wqc/national-recommended-water-quality-criteria-human-health-criteria-table
[25] https://www.epa.gov/wqc/national-recommended-water-quality-criteria-aquatic-life-criteria-table
[26] https://www.epa.gov/wqc/national-recommended-water-quality-criteria-organoleptic-effects
[27] https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations
[28] https://www.epa.gov/sdwa/secondary-drinking-water-standards-guidance-nuisance-chemicals

measured parameters, data provider names and classification, and monitoring frequency categorization. A count of the number of monitoring instances (that is, distinct days with data within the two-year period of interest) during the period of interest at each site was also added.

These site data were merged with site metadata to generate a CSV with geospatial coordinates for use in the dashboard.

WQP Data update

In the update script, metadata are retrieved for all sites for which data has been collected since the last update that were not already represented in the sites data file. These are combined with the previously-accessed metadata. All data from the period of record for all the monitoring sites is then pulled and analyzed as above, as marginal additions would require retaining all data regarding what parameters were included in each monitoring instance.

The update script also handles data enhancement for site records, adding classifications for sites based on whether they pertain to surface water, groundwater, or other water sources and identifying for each site the EPA region within which it is located. State FIPS codes are also converted into state abbreviations and names to facilitate data presentation.

The combined site metadata and monitoring instance data are saved as a CSV with geospatial coordinates at the end of the update script.

ACS Demographic Data

ACS Metadata

Geospatial boundaries and characteristics for Census geographies at the block group and tract scale were accessed using the 'tidycensus' package for R.

Census Geography Characteristics

The population and demographic variables selected for inclusion and analysis in the dashboard included income, race, and ethnicity. Other data contained within the ACS, related to characteristics such as sex, poverty, language, and home ownership (tenure), could provide the basis for fruitful further analyses, but race, ethnicity, and income were prioritized for the present analysis because of their assumed connection with questions of environmental justice and equity.

Data pertaining to racial and income demographics are available at the block group scale, while data pertaining to ethnic demographics are only available at the tract scale. These variables were joined to the relevant census geographies. Income data were reclassified to create equal-interval household income categories.

ACS Data Update

ACS data are not updated by the update script as new ACS data are only released once a year. On an annual basis as new ACS surveys are released, the initial script can be updated with the most recent survey end year and rerun to update the demographic data for the subwatersheds.

HUC12 SUBWATERSHEDS

Subwatershed Metadata

Metadata for the HUC12 subwatersheds were accessed using the 'nhdplusTools' package for R. HUC12 metadata included information about the name, type, intersected states, and area of each subwatershed, as well as its geospatial extent. Some subwatersheds had multiple records within the dataset due to multipart polygon features. The dataset was consolidated so that each of the 94577 distinct subwatersheds was represented by a single record. Geographies from the Watershed Boundary Dataset accessed through 'nhdplusTools' are provided in the NAD83 coordinate reference system.

Subwatershed Characteristics

Since subwatersheds are the primary unit of analysis for this consideration, demographic data provided by census geographies (block group and tract levels) were transferred to these hydrologic geographies using areal-weighted interpolation. This method assigns values to each HUC12 in proportion to the composition of the census geographies with it contains or intersects. Areal-weighted interpolation assumes homogenous spatial distribution of the demographic characteristics reported for each block group or tract, an assumption that may not accord with reality. However, this assumption is generally understood and accepted as standard in geospatial use of census data. All of the demographic data in consideration are spatially extensive (that is, they are expected to increase with area, like population or number of households); therefore, this aggregation computed the sum of all persons or households in each variable category of the block groups or tracts in proportion to their spatial intersection with the subwatershed.

The subwatersheds were also classified according to the EPA region to which each belongs, allowing for both state (included in the original dataset) and regional analyses of trend in monitoring and data provision.

Monitoring sites were considered to provide monitoring coverage for the subwatershed (HUC12) in which they are located as well as for any subwatershed to which it directly drains. Stated otherwise, a HUC12 is considered to be served by monitoring if it contains at least one monitoring site from the WQP sites dataset as derived above, or if it is immediately downstream of a HUC12 containing a monitoring site. Therefore, the monitoring sites attributed to each HUC12 were determined by calculating the number of sites occurring within the subwatershed ("sites within") and adding the number of sites in any subwatershed from which it drains ("sites upstream"). As both the site points and the subwatershed polygons are in the NAD83 coordinate reference system, no transformation was needed to ensure compatibility prior to intersection. The sites within each subwatershed were determined by counting the number of WQP monitoring sites that spatially intersect the HUC12. The sites upstream were then determined in reliance on the downstream drainage relationships communicated by the "tohuc" component of the NHDPlus from which HUC12 metadata were obtained. These counts were added together to calculate the total number of sites by which water quality for each HUC12 is monitored.

The HUC12 data are then saved as an ESRI shapefile.

Subwatershed Data Update

The metadata and demographic data of the subwatersheds only need to be updated annually when new ACS products are released by the census bureau, as the geographies of the WBD are

relatively fixed. However, the number of sites considered to monitor a given subwatershed will change as new sites and data are added. Accordingly, calculations of the sites within and sites upstream of a given subwatershed are rerun in the update script. The update script also classifies subwatersheds as “monitored” or “unmonitored” and converts state FIPS codes into state abbreviations and names to facilitate data presentation.

The updated HUC12 data are re-saved as an ESRI shapefile at the end of the update script.