Supporting Information

Allaire et al. 10.1073/pnas.1719805115

SI Literature Review

Few peer-reviewed studies have addressed water system compliance with the Safe Drinking Water Act (SDWA). Only two national assessments could be found in the peer-reviewed literature (e.g., refs. 1 and 2). Rubin (1) describes summary statistics of violations for four SDWA rules for the year 2011. This analysis does not include community demographics and only addresses four SDWA rules—Total Coliform, disinfection byproducts (DBPs), Arsenic, and Lead and Copper. A major limitation is that since this paper presents only summary statistics, the results do not isolate the association between violations and specific utility characteristics. Rubin (1) finds that similar proportions of small and large systems incur health-based violations. However, given that only summary statistics are presented, results do not isolate the association between violations and specific utility characteristics.

Meanwhile, Wallsten and Kosec (2) develop count regressions that relate community and water system characteristics to SDWA violations over a 7-y period (1997–2003). The analysis includes every public water system (PWS), not only community water systems (CWSs). In addition, demographic variables are considered to be time invariant. Wallsten and Kosec (2) find that private ownership does not generally affect compliance. Among smaller systems, private ownership is associated with fewer maximum contaminant level (MCL) violations, yet greater monitoring and reporting violations, which raises the question of whether fewer reported MCL violations are driven by inadequate monitoring and reporting.

Other past studies generally include limited geographic areas. Several studies focus on a single state or province (e.g., refs. 3–7). Pike (6) included only five water systems in Pennsylvania in a probabilistic Bayesian network model to assess how operator decisions and system characteristics influence SDWA violations. Results indicate that the likelihood of violation is particularly associated with operator decisions, such as filter backwash frequency. Coulibaly and Rodriguez (7) use 10 Quebec utilities to analyze how different disinfection techniques influence total coliform violations. The analysis was limited to comparison of means and ordinary least-squares regression. Results indicate that total coliform violations are associated with lower chlorine residuals. Rahman et al. (4) focused on water systems in Arizona in a cross-sectional regression analysis of MCL violations. The likelihood of MCL standards is found to increase with larger utility size and public ownership.

Only 327 CWSs in central California were included in the hierarchical linear models used by Balazs et al. (3) to analyze the association between nitrate levels in drinking water and demographics at the census block group level. This study finds that utilities serving larger proportions of Latino residents tend to have higher nitrate concentrations. Last, Guerrero-Preston et al. (5) focused on a small sample of CWSs (n = 239) over a 5-y period in logistic regression analysis that assessed factors associated with SDWA compliance of small, rural CWSs in environmental justice communities in Puerto Rico. The vast majority of small, rural systems in Puerto Rico were noncompliant with the SDWA during the study period. This study found that noncompliance is less prevalent among systems that had installed treatment equipment.

SI Data

We assembled a panel dataset that includes 17,900 CWSs in the United States from 1982 to 2015. Inclusion criteria were restricting the study sample to CWSs that began reporting violations in 1982 or prior, serve over 500 people, and are located within the continental United States. CWSs serve year-round populations and are subject to all SDWA regulations. This is in contrast to noncommunity systems that serve transient populations (e.g., gas stations, campgrounds), which are exempt from most drinking water regulations, except for contaminants that pose an immediate health risk.

Water systems included in the analysis are those that report to the EPA Safe Drinking Water Information System (SDWIS) every year of the study period. By considering the reporting start date, we limit our sample to CWSs with consistent reporting over the study period. Water systems can begin reporting in years later than 1982 for a variety of reasons, including establishment of a new water system and a preexisting system growing in size to become subject to SDWA regulation. In addition, we only include CWSs serving more than 500 people because SDWA regulations apply to small CWSs differently, especially in terms of sampling frequency, which could influence the likelihood of detecting a violation. For example, some DBPs only need to be sampled on an annual basis, rather than quarterly, for CWSs serving less than 500 people. Furthermore, very small systems are more likely to have inadequate reporting practices (8).

Very small systems, serving fewer than 500 people, do not have significantly different rates of violation occurrence for most contaminant types. Rubin (1) also provides evidence that very small CWSs do not have a disproportionate number of violations. Therefore, we do not believe that our sample excludes the worst violators. Moreover, very small systems only serve 1.5% of the US population. Thus, excluding these CWSs is not expected to considerably change the generalizability of our findings.

A limitation of the SDWA violations dataset is underreporting. The EPA estimates that 38% of health-based violations were not reported by states to the SDWIS from 2002 to 2004 (8). An audit by the Government Accountability Office in 2009 revealed that 26% of health-based violations were either not reported or inaccurately reported (9). Furthermore, reporting quality can vary across contaminate types and states. For example, it is estimated that 68% of national violations of the total coliform rule and only 15% of other MCL violations are reported accurately (8). Data quality of health-based violations does appear to be improving over time. During 1996–1998, only 40% of health-based violations were reported in SDWIS; by 1999–2001 this improved to 65% (10). Reasons for inaccurate reporting include insufficient training, staffing, and funding at the state level to conduct enforcement activities.

The study sample represents about 36% (n = 17,900) of the 50,121 CWSs in the full SDWIS dataset. In the full dataset, 34,242 CWSs reported violations throughout the 1982–2015 study period and are located in the continental United States. Of these, 18,461 systems served populations of more than 500. A further 561 systems were dropped due to incomplete reporting of CWS characteristics. This produces our final sample of 17,900 water systems. In terms of health-based violations, the final sample contains 95,754 records, for CWSs that meet our inclusion criteria. The full EPA SDWIS dataset contains 343,119 records of health-based violations at all PWSs from 1979 to 2015. The full dataset also has 220,930 records of health-based violations at CWSs from 1979 to 2015.

Health-Based Violations. The EPA SDWIS provides information about PWSs and their violations of drinking water regulations, as reported to EPA by each state. Our study focuses on health-based

violations, which include MCLs, maximum residual disinfectant levels, and treatment techniques. Standards for contaminants that can trigger health-based violations are specified by the National Primary Drinking Water Regulations. MCLs specify the highest allowable concentrations of contaminants in drinking water. MCLs represent the threshold below which health risks are minimal. Maximum residual disinfectant levels specify the highest concentrations of residual disinfectants in drinking water systems, to limit the risk of exposure to DBPs. Last, treatment techniques require certain processes that are intended to reduce contaminant levels.

We define the year in which the violation occurred as the year of the compliance period start date. We assume that if a CWS does not have a violation entry for a given year, then no violation occurred. However, it must be noted that a missing violation entry does not necessarily mean no violations occurred; it only signifies that no violations were reported.

We define the following violation categories.

Total coliform. Total coliforms include a variety of bacteria that are mostly not harmful to humans, but serve as an indicator for *E. coli*, parasites (e.g., *Cryptosporidium*, *Giardia lamblia*), and viruses. Harmful microbiological contaminants that can be indicated by total coliforms are usually associated with gastrointestinal illness. Some types of bacteria, such as *E. coli*, can cause acute gastroenteritis, which can be fatal for vulnerable individuals. Violations can be classified as acute or nonacute, where acute violations involve positive detection of fecal coliform. Our total coliform includes violations of the Total Coliform Rule.

Treatment rules and nitrate. This category includes nitrate, nitrite, and the various contaminants regulated under the Surface Water Treatment Rules and Ground Water Rule. Nitrates and nitrites can interfere with the ability of blood to carry oxygen and can pose an acute health threat. There is some evidence of a relationship between nitrates and stillbirth, "blue-baby" syndrome, and low birth weight. As a result, nitrate MCL violations require public notification within 24 h.

Arsenic. Arsenic is a naturally occurring element, and long-term exposure can cause bladder and lung cancer, as well as cardiovascular and neurological disorders. In children, links have been shown between arsenic and low birth weight (11). Generally, groundwater systems in the United States have higher levels of arsenic than surface water supplies (12). The MCL for arsenic was set at 50 μg/L in 1975 and was revised to 10 μg/L in 2001, with final implementation by 2006.

Lead and copper. Lead and copper enter drinking water mostly from corrosion of service lines or in-home plumbing that contain lead or copper. Lead can cause damage to the brain, red blood cells, and kidneys. Health risks associated with copper exposure include liver and kidney damage. No MCL exists for lead or copper. Instead, there is an "action level," which is based on the 90th percentile level of tap water samples. Exceedance of an action level is not considered a MCL violation but can trigger actions such as corrosion control treatment, additional monitoring, and lead service line replacement. The Lead and Copper Rule was enforceable in 1992 and was revised in 2007 to improve implementation of monitoring, treatment, and customer awareness.

Other violations. "Other" contaminants include DBPs, radionuclides, and organic and inorganic chemicals. Many of these contaminants are associated with chronic health effects due to long-term exposure. DBPs are formed due to reactions between disinfectants and naturally occurring materials in water. Long-term exposure is associated with cancer or nervous system problems. Meanwhile, radionuclides are also probable carcinogens and can enter drinking water from natural sources or rare releases from laboratories or nuclear power plants. Organic and inorganic chemicals are associated with a variety of long-term health effects including cancer and kidney damage.

Water System Characteristics. Water system characteristics include type of source water, service population, and ownership type. These are time invariant since the SDWIS provides only the most recent year of water system characteristics. Water source is classified based on primary source and includes the categories of purchased water, surface water, and groundwater. It should be noted that systems often have several sources, yet the SDWIS only publicly reports a system-wide classification. Purchased water is obtained from other utilities and most commonly originates from surface sources (84% of purchased water originates from surface water). Groundwater systems rely on wells and can be under the influence of surface water. Meanwhile, if a system relies on any surface water source, it is classified as a surface water system (13).

Utility size categories are defined based on population served by a given utility. The EPA SDWIS provides local retail population served for each water system. The population count does not include customers of another CWS that purchases water from a given system. Our size categories follow EPA designations—small utilities serve 501–3,300 people, medium serve 3,301–10,000, and large serve more than 10,000 (14).

Ownership of utilities is categorized as public or private. Public ownership includes government (federal, state, or local) and Native American tribes. We exclude utilities without a known ownership type or that are classified as "public/private" since this category does not have a consistent definition (2).

We create a Herfindahl–Hirschman index (HHI), which measures the ownership concentration in the local water market. It is calculated as the sum of squared market share of each firm. In the case of water utilities, we define a county as being the local water market. The share of total water accounts in a county that a utility serves is its market share. If a market is highly competitive and composed of a large number of firms of relatively equal size, the HHI will approach zero. Meanwhile, HHI will increase as the number of firms in a market decreases and as differences in size increase between firms. We scale HHI values from 0 to 1 for use in our regressions. Greater HHI values indicate fewer systems and greater differences in size; this makes it challenging for regulators to benchmark the performance utilities in a given location.

County-level locations of each CWS are defined based on the CWS address provided in the SDWIS. In this study, the county-level location (identified by FIPS code) is used to merge the SDWIS violation records with census data. We assign a county FIPS code (2010 census definition of FIPS codes) by obtaining the geographic coordinates of the CWS address in ArcGIS and then using the TIGER/Line Shapefile, which contains the detailed county boundary information (15).

Community Characteristics. Community characteristics are county-level US census variables and represent annual values for each year of the study period. The decennial census data (1970, 1980, 1990, 2000, 2010) and the American Community Survey data in the year of 2006, 2008, and 2012 were obtained via Social Explorer. Assigning census information of one county to each CWS is reasonable, given that over 97% of systems in our study serve only a single county.

Data availability was sufficient to interpolate values for inter-census years. All annual data from 1980 to 2011 are interpolated by the method of monotone piecewise cubic interpolation (16). Projections of variables were made from 2013 to 2015. Cubic splines interpolate data with piecewise cubic polynomials. The curves interpolated by the monotone cubic spline are relatively smooth. Other interpolation methods were also tested, such as natural splines, quadratic splines, and the Forsythe, Malcolm, and Moler's method. Household income is in terms of 2015 US dollars, adjusted by consumer price indexes.

The variable percent nonwhite population is calculated as 1 minus the percentage of white population. Housing density is

included in the regression analysis and housing density categories are defined for rural [<16 units per square mile (sq. mi.)], suburban (16–380.7 units per sq. mi.), and urban (>380.7 units per sq. mi.), based on categorizations from ref. 17.

SI Materials and Methods

Hot-Spot Analysis of Violations. Spatial trends and hot spots are assessed via local spatial autocorrelation. Individual counties of local clusters are identified by determining the spatial dependence and relative magnitude between a given county and neighboring counties. Specifically, we use the local $Gi^*(d)$ statistic (local Getis–Ord statistic) (18). The greater the magnitude of the statistic, the more intense the clustering. A value close to zero indicates no spatial clustering. To be a statistically significant hot spot, a county must have a large number of violations per water system and be surrounded by counties that also have large numbers of violations. The local Getis–Ord statistic can be written as follows:

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{i,j}(d)x_j - W_i\bar{x}}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - W_i^2}{n-1}}}, \qquad \text{[S1]}$$

where x_j is a measure of violation occurrence in county j, $w_{i,j}$ is the spatial weight between county i and j, W_i is the sum of weights $w_{i,j}$, $\bar{x} = 1/n\sum_{j=1}^{n} x_j$, and $s = \sqrt{1/n\sum_{j=1}^{n} x_j^2 - \bar{x}^2}$.

Two sets of local Getis–Ord statistics are estimated, one set where x_j is the number of violations per CWS in county j. The second set represents x_j as the number of CWS-year observations with at least one violation per CWS in county j. The spatial weights ($w_{i,j}$) were defined as equal to 1 if $d_{ij} < d$, and zero otherwise, where d_{ij} is the Euclidean distance between centroids of county i and j. The threshold distance (d) was selected such that each county had at least one neighbor, which was a distance of 146.2 km.

Probit Regression Models. Probit regression is used to assess the relationship between quality violations characteristics of water systems and communities. Covariates included in the probit models were selected based on a careful review of the literature, least absolute shrinkage and selection operator (lasso) regression, and specification tests. The final model specifications were based on lowest values of Akaike information criterion and Bayesian information criterion. Our preferred model specification is presented in the main text (Eq. **1**). Models are specified for total violations and for total coliform violations. Total coliform violations are assessed separately to determine whether utility and community characteristics influence this class of contaminant differently.

We estimate coefficient values and average marginal effects. Average marginal effects are useful for interpretation; they provide a single estimate of the effect of each covariate on $\Pr(Y = 1)$. To calculate average marginal effects, two hypothetical populations are compared. Each observation is treated as though it had the characteristic of interest (e.g., private ownership = 1). We compute the probability that this observation would have a violation. Then, we compute the probability for the case where the observation does not have the characteristic of interest (e.g., private ownership = 0). The difference in the two probabilities in the marginal effect for that observation. Then, the average of all marginal effects across all observations is computed, which results in the average marginal effect (AME).

Variable Selection for Probit Models.

Lasso regression and specification tests. Due to the large number of potential covariates and concerns of multicollinearity, lasso regression provided a way to improve selection of appropriate independent variables. It is a shrinkage and selection method for regression models (19). This technique penalizes variables that do not improve fit. The penalty term (λ) modifies the original coefficient estimated with binary regression. As the penalty term grows, the coefficient estimate of the independent variable will approach zero. Alternatively, as the penalty decreases, the coefficient estimate will resemble the original regression. The lasso is a constraint on the sum of the absolute values of the model parameters (β_m), excluding the intercept (β_0). This constraint can be expressed as follows:

$$\sum_{m=1}^{p} |\beta_m| \le \lambda; \quad \lambda > 0. \qquad \text{[S2]}$$

In lasso regression, one objective is to select the smallest value of λ after which the coefficient estimates stabilize. For example, in the lasso regression for total coliform violations, after using a cross-validation procedure, we selected a very small λ value of 0.00028.

Results of the lasso regression indicated that annual climate variables, fertilizer use, and age of housing stock do not improve model fit. Therefore, these variables were not included in the final model specifications. The covariates that were especially informative were a lagged violation indicator, primary water source, median household income, and utility size indicators. Some census variables, such as education and marital status, were excluded from final specification due to multicollinearity concerns. Education was highly correlated with household income, while marital status was highly correlated with percent nonwhite population.

Higher-ordered terms were included in the probit models based on specification tests, such as the Pregibon's link test. In the final specifications, a link test indicated that higher-order terms are unlikely to be missing from the model since the square of the prediction is insignificant (value of $P > 0.05$ for all specifications). This also suggests that the probit specification is appropriate.

Literature review for variable selection. Selection of covariates also relied on theory, based on a careful review of the literature. An indicator of lagged violation occurrence was included since noncompliant CWSs tend to have reoccurring violations. In our regression sample, nearly 74% of CWSs had at least 1 y with a repeat violation. Persistence could be due to a lack of ability to upgrade water treatment facilities (20).

Water source is also anticipated to be associated with violation occurrence since different types of water sources may be more prone to contamination. Purchased water sources might be less prone to SDWA violations since wholesale water providers may have greater capacity to achieve regulatory compliance (2). Different water sources are likely prone to different types of contaminants. For example, surface water generally contains more organic material, while groundwater can contain naturally occurring contaminants, such as heavy metals (21). Groundwater can also be especially vulnerable to inappropriate wastewater disposal and agrichemical pollutants (22). Water source will also influence the types of regulatory requirements to which a utility is subject. Groundwater systems face fewer regulatory requirements, compared with surface sources.

Utility size is also associated with the types and stringency of regulations to which a utility is subject, as well as capacity to comply with quality standards. The sign of the correlation between large utilities and violation occurrence is ambiguous. A negative association is possible if larger and more extensive treatment and distribution systems are susceptible to more potential problems. In addition, large utilities have greater capacity for regulatory compliance. Generally, smaller utilities have less technical and financial capacity to maintain their systems and

meet regulatory guidelines (9). Negative associations could also be attributable to differences in technology requirements by utility size (23). Smaller utilities tend to have treatment systems that are not as equipped as larger utilities to effectively treat a broad range of contaminants, especially newly regulated contaminants. For example, treatment techniques for small utilities tend to be limited to simple chlorination. However, smaller water systems (service population fewer than 3,301) have been found to be no more likely to have health-based violations, compared with larger systems, except very large systems (service population over 100,000) (1).

Alternatively, larger systems might be susceptible to a wider range of potential problems due to more extensive treatment and distribution systems. In addition, greater regulation faced by larger utilities could cause the association to be positive since violations might be more likely to be detected in large systems. Larger utilities tend to be more tightly regulated in terms of sampling and reporting. A greater number of water samples taken at more sampling locations are typically required for larger systems (24). In addition, large systems (>10,000 people served) must send an annual water quality report to customers, which has been shown to increase compliance (25). Larger firms in other contexts have been found to be more likely to overcomply with environmental regulations (26, 27). Greater sampling could imply a higher likelihood of detecting a violation at larger utilities for reasons that are unrelated to actual water quality.

Type of ownership, private or public, might influence water system performance and finances. Ownership and utility sector performance is debated in the literature, but there is no clear consensus (28). Some studies find no difference in regulatory compliance between privately and publicly owned utilities in France (29) and in the United States (2). Among smaller water systems, private ownership might be associated with fewer MCL violations (2). However, other studies find that larger and publicly owned PWSs might be more likely to violate a MCL (4). Private systems could be less likely to incur violations due to differences in regulation and incentives. Private firms face the possibility of elimination, which could lead to greater compliance (30). In addition, while private and public utilities face the same level of oversight for water quality, oversight of utility finances can differ. For example, in some states private utilities are subject to stricter regulation of water rates and financial reporting, which could incentivize improved system performance (31).

Alternatively, local government systems might be subject to more immediate accountability to their customers since expenditures tend to be controlled by elected local officials. In addition, local government systems had to have greater access to subsidized financing. This could lead to government-owned utilities incurring fewer violations. However, differences across government types likely exist. For example, water systems owned by Native American tribes have evidence of higher violation occurrence (14).

The HHI is intended to be an indicator of benchmark competition. Water suppliers in the United States do not compete for customers since service areas do not overlap and customers do not choose their water supplier. Instead, the HHI provides a measure of the extent to which utility performance might be compared across suppliers by a regulatory agency. As the market becomes more concentrated, there is less opportunity to benchmark the performance of any one utility against another. As ownership concentration increases, fewer benchmark comparisons are possible, and this might reduce a water supplier's incentive for high performance in terms of regulatory compliance (2).

Interactions of utility characteristics are also important to consider. For example, smaller systems are more likely to be privately owned as well as rely on groundwater sources, based on our summary statistics. Therefore, comparing differences in violation occurrence across only one utility characteristic might be misleading. The final model specifications include a variety of interaction terms for system size, ownership, and source water.

Community characteristics included median household income, housing density, and percent nonwhite population. It is anticipated that communities with higher income will have the financial resources to allow their water systems to be less likely to incur quality violations. Previous studies have found that areas with greater poverty levels have higher numbers of SDWA violations (32). Housing density is expected to have a negative association since utilities serving urban areas and larger populations might be less likely to violate. In rural areas, small systems in particular face difficulties because of declining populations and lower incomes (33).

Few studies in the United States look into disproportionate exposures to drinking water violations by low-income and minority communities. Studies in California find that water systems serving larger percentages of Latino residents tend to have higher concentrations of nitrate (3) and higher violation counts (32). A more recent study uses negative-binomial regression models to assess association between counts of SDWA violations and socioeconomic status as well as percent Hispanic population for the United States in 2010–2013 (34). One limitation of this analysis is the possibility bias due to omitting an indicator of housing density.

SI Results

Summary Statistics. Summary statistics and definitions for all variables included in the statistical analysis are provided in Table S1. Our balanced panel dataset contains 34 y and 17,900 CWSs, which serve 87% of the population supplied by CWSs in the continental United States. Our study sample includes 48% of health-based violations (95,754 out of 198,418) and 36% of CWSs and in the continental United States from 1982 to 2015. The portion of the US population served by noncompliant systems fluctuates year to year. On average, 19 million people are served by systems that incur a health-based water quality violation.

Notably, the relative proportion of some violation types has also changed over time. Total coliform was the most common type of violation until 2003, when "other" violations became more prevalent. Other violations dramatically increased from just 554 violations in 2002 to 3,581 in 2005. This rise is likely attributable to a series of new regulations for DBPs and radionuclides.

Utility characteristics. In our study sample, small utilities represent more than one-half (56%) of systems, while the remaining utilities are nearly evenly split between medium-sized (23%) and large systems serving 10,000 or more people (21%). Therefore, CWSs are highly fragmented, with the vast majority serving small populations. This arrangement differs dramatically from the natural gas and electricity sectors in the United States, which are composed of about 1,200 and 3,000 companies, respectively (35).

Since very small CWSs (serving 500 people or less) are excluded from the analysis, the regression sample contains a greater proportion of medium and large systems than the full SDWIS database. Therefore, our results cannot be generalized to systems serving 500 people or less. In terms of violation counts, the portion of violation counts borne by each size class tends to follow the portion of CWSs within a given size class. For example, small utilities represent 56% of CWSs in the sample, and they incur 52% of the total number of violations. Similarly, larger utilities (serving >10,000 people) represent 21% of CWSs and bear 23% of the violations.

Groundwater is the most prevalent water source (58% of CWSs), followed by purchased water (26% of CWSs) and surface water (16% of CWSs).

Private ownership was uncommon, with only 14% of utilities being privately owned. In terms of violation counts, private utilities incurred 10% of total violations (9,836 out of 95,754 total), which is slightly lower than the portion of private utilities in the overall

sample of CWSs (14%). Local government ownership is the most common classification (85% of CWSs), while other types of government ownership are the least common (2%). Private ownership is more common among smaller utilities. About 17% of small utilities are privately owned, compared with 9.6% of utilities serving more than 3,300 people. Thus, utility size and ownership are related and the final regression specifications control for interactions between size and ownership. For example, an interaction term for private ownership and large utility size is created.

Interaction terms between ownership, system size, and source water were included to determine whether associations with violation occurrence differ between CWSs that have different sets of characteristics. Additional terms were included to account for the relationship between utility size and source water. Larger utilities are more likely to have a surface water source. Nearly 17% of CWSs in the sample rely on surface water, and these are mostly large systems (7% of regression sample) or medium-sized (4.6% of regression sample). Last, since many private systems are small, this means that private CWSs are less likely to rely on surface water. Therefore, an interaction was created for private CWSs and surface water source.

Community characteristics. Community characteristics vary substantially across CWSs and years in the regression sample. Median household income ranged from $13,101 (in 2015$) to $125,624, with a mean of $52,211. Large variations also existed for percent nonwhite population (0–87%) and housing density (0.2–6,009 housing units per square mile). The mean nonwhite population is 15%, which is less than the overall nonwhite population in the United States from 1980 (17%) to 2015 (23%) (36, 37). This is likely attributable to larger utilities tending to serve more diverse, urban areas. The mean value of housing density (146 units per sq. mi.) is considerably greater than nationwide values, which increased from 25 to 37 units per sq. mi. from 1980 to 2010 (38, 39). This is to be expected since the analysis only includes populations that are served by a CWS supplying services to more than 500 people. Therefore, extremely small communities and housing developments are excluded from the analysis.

The HHI measures the concentration of firms in an industry. The US Department of Justice considers a market to be highly concentrated if the HHI value exceeds 2,500, equivalent to 0.25 on a 0–1 scale (40). As might be expected for the water industry, the values of HHI tend to be quite high in our regression sample (mean HHI is 0.316). However, HHI values range from 0.03 to 1, indicating that some areas might have moderately concentrated or even not concentrated markets.

Several covariates are moderately correlated. This is expected since rural areas tend to be served by smaller systems and populations are lower income and less diverse. Log of housing density is moderately correlated with large systems ($r = 0.35$), purchased water source ($r = 0.21$), log of median income ($r = 0.57$), and percent nonwhite population ($r = 0.24$). In addition, large systems are moderately correlated with surface water source ($r = 0.23$) and log of median income ($r = 0.23$).

Violation occurrence: Differences across CWSs and community characteristics. Differences in the rate of violation occurrence exist across characteristics of CWSs and communities (Table S2). Privately owned utilities have slightly lower violation occurrence than those owned by local governments. Violations occur in only 6.5% of utility-year observations for private utilities, compared with a greater portion of observations with violations in local government (8.3%). Meanwhile, utilities that purchase water have fewer observations with violations than utilities that rely on groundwater, both for total coliform and total violations (Table S2). Purchased water is treated water obtained from other utilities; most commonly, it originates from surface sources. While 26% of CWSs rely on purchased water, the most common water source is groundwater (58% of CWSs rely on groundwater) (Table S1).

In addition, we find that CWSs in rural counties tend to have higher rates of violation occurrence than in urban counties. The rate of total coliform violations for CWSs in rural counties is 56% higher than for urban counties. Meanwhile, the rate of all violation types in rural counties is 76% higher. In rural counties, 9.5% of utility-year observations have violations, compared with 5.4% of observations in urban counties (Table S2). Differences in incidence of DBP violations are especially large—incidence is more than three times higher in rural areas than urban areas. The vast majority of CWSs are located in suburban counties (60% of CWSs in our sample), while few are located in urban counties (8%). Categories of urban, suburban, and rural areas are based on housing density, as described in *SI Data*.

Spatial Trends. Violations also vary considerably across geographic locations. Both quality of source water and state-level enforcement can influence these differences across counties. Source water quality can vary due to naturally occurring contaminants or anthropogenic impacts. Meanwhile, state-level enforcement can differ due to variation in sampling protocols, technical capacity, and financial resources.

Differences in violations across states are similar to differences across counties (presented in the main text, Fig. 4). At the state-level, in terms of total violations per CWS, Oklahoma, District of Columbia, Idaho, and Nebraska have highest rate of violation incidence. When we focus on extremely poorly performing counties, these same states continue to bear the greatest incidence. States with more than 20% of counties in the top 10th percentile of violation incidence per CWS include Oklahoma, Idaho, Nebraska, Kentucky, Kansas, Oregon, and District of Columbia.

Across the full study period, the Southwest had a mean violation incidence of 8.6 violations per CWS, which is much greater than mean values in the Plains region (4.8 violations per CWS) and Great Lakes (3.8). About 73% of violations in the Southwest were due to DBPs in 2005. The sharp rise in DBPs likely reflects the challenge in complying with new rules.

A large portion of water systems in the Southwest region have repeat violations in subsequent years. Oklahoma in particular has frequent noncompliance, given that 13% of CWSs in the state have incurred violations in 8 subsequent years or longer. This proportion far exceeds other states by an order of magnitude.

Fig. S1. Health-based violations by region and year.

Fig. S2. Spatial clusters (hot spots) of health-based violations, by county and time period.

Regression Results. Table S3. Regression results: Coefficient estimates.

Table S4. Regression results: Marginal effects.

Fig. S3. Average marginal effects (AMEs) of year dummy variables.

Table S5. Time trend regression results: Coefficient estimates.

Fig. S4. Deviation of state from national time trend.

Average adjusted predictions. The highest predicted probability of violation is estimated to occur at small, rural CWSs relying on surface water sources. To assess how associations differ across values of covariates, we calculate average adjusted predictions. This involves calculating the probability of violation for a water system with specified characteristics. We present results for how the probability of violation varies across values of housing density, water source, and system size. Across values of housing density, significant differences exist in likelihood of violation between primary water sources and utility size (Fig. S5).

Fig. S5. Average adjusted predictions of water source and system size, across values of housing density.

Principal-component regression. We conduct principal-component regression to better understand the relationship between violation occurrence and three covariates that are moderately

correlated: housing density, median household income, and percent nonwhite population. In the probit regression results, it is difficult to interpret the coefficient estimates of these covariates. Therefore, we first conduct principal-component analysis (PCA) and then use the principal components (PCs) as covariates in alternative specifications of the probit models. Since the PCs are uncorrelated, there is no multicollinearity between our original three variables.

Principal-component regression involves first conducting PCA and then using the PCs as covariates. PCA is a dimension-reduction method that identifies a linear combination of variables such that a large portion of the variance is accounted for. PCA transforms correlated variables into a smaller number of uncorrelated variables called PCs. The first PC accounts for as much of the variance as possible, and each succeeding component accounts for as much of the remaining variability as possible. PCA analyzes total variance, which includes both common and unique variance.

First, we conduct PCA on the correlation matrix of the three variables of interest. Since these variables have difference units of measurement, standardizing these data are crucial. We retain the PCs that have an eigenvalue greater than 1. This results in retaining the first two PCs,, which account for 89% of the total variation in the three variables. The first PC has an eigenvalue of 1.58 and explains 53% of the variation in the three variables, while the second PC has an eigenvalue of 1.09 and explains 36% of the variation.

Unrotated component loadings for these PCs are as follows. The first PC represents that common variance of all three correlated variables. The first PC has component loadings of 0.723 ln (Housing Density), 0.673 ln(Median Household Income), and 0.159% nonwhite population. The second PC represents the unique variance, not included in the first PC. The first PC has component loadings of 0.141 ln(Housing Density), −0.369 ln (Median Household Income), and 0.919% nonwhite population. The communality of each of the three variables are as follows: 0.542 ln(Housing Density), 0.589 ln(Median Household Income), and 0.869% nonwhite population. Communality is the total influence on a single variable from all of the PCs associated with it. It is equal to the sum of all of the squared component loadings for all PCs related to the variable; a value of 1 indicates that the variable can be fully defined by the PCs.

A positive relationship exists for all coefficients of the first component, with the largest contributions associated with housing density and income. This component can be interpreted as a general indicator of urbanization. More urban areas are typically associated with dense housing, higher incomes, and greater racial diversity. The second PC is dominated by nonwhite population, and to a lesser extent, is associated positively with density and negatively with income. This component indicates the variability among water systems that reflects greater nonwhite population and housing density, but lower household incomes. Based on the component loadings for the first and second PCs, we calculate component scores for each observation in our sample. These component scores are then included in alternative specifications of the probit models.

PC regression results indicate that the two PCs are significantly associated with violation occurrence. The coefficient estimate of the first PC represents the average effect of all three variables, although housing density and income have the largest component loadings. Meanwhile, the coefficient estimate of the second PC represents the contrasting effect. The coefficient estimates of the two PCs are jointly significant in both the total violations and total coliform models.

The first PC is negatively associated with violation occurrence, both any violations and total coliform. This can be interpreted as increased urbanization, mostly dense housing and higher income, being associated with reduced likelihood of violations. The marginal effect of the first PC on total coliform violations is equal to the marginal effect estimated in our original total coliform model, model 4 (Table S6). This suggests that effect of housing density dominates the coefficient estimate of the first PC. The second PC is positively associated with total coliform violation occurrence but is not significantly associated with total violation occurrence. This might indicate that greater nonwhite population and lower income are associated with higher likelihood of total coliform violations.

1. Rubin S (2013) Evaluating violations of drinking water regulations. *J Am Water Resour Assoc* 105:E137–E147.
2. Wallsten S, Kosec K (2008) The effects of ownership and benchmark competition: An empirical analysis of U.S. water systems. *Int J Ind Organ* 26:186–205.
3. Balazs C, Morello-Frosch R, Hubbard A, Ray I (2011) Social disparities in nitrate-contaminated drinking water in California's San Joaquin Valley. *Environ Health Perspect* 119:1272–1278, and erratum (2011) 119:A509.
4. Rahman T, Kohli M, Megdal S, Aradhyula S, Moxley J (2010) Determinants of environmental noncompliance by public water systems. *Contemp Econ Policy* 28:264–274.
5. Guerrero-Preston R, Norat J, Rodríguez M, Santiago L, Suárez E (2008) Determinants of compliance with drinking water standards in rural Puerto Rico between 1996 and 2000: A multilevel approach. *P R Health Sci J* 27:229–235.
6. Pike W (2004) Modeling drinking water quality violations with Bayesian networks. *J Am Water Resour Assoc* 40:1563–1578.
7. Coulibaly H, Rodriguez M (2003) Spatial and temporal variation of drinking water quality in ten small Quebec utilities. *J Environ Eng Sci* 2:47–61.
8. US Environmental Protection Agency (2002) Data reliability analysis of the EPA safe drinking water information system/federal version (SDWIS/FED) (Office of Water, US Environmental Protection Agency, Washington, DC), EPA 816-R-00-020.
9. Government Accountability Office (1990) Drinking water: Compliance problems undermine EPA program as new challenges emerge (Government Accountability Office, Washington, DC), GAO/RCED-90-127.
10. US Environmental Protection Agency (2004) EPA claims to meet drinking water goals despite persistent data quality shortcomings (US Environmental Protection Agency, Washington, DC), Report No. 2004-P-0008.
11. Rahman A, et al. (2009) Arsenic exposure during pregnancy and size at birth: A prospective cohort study in Bangladesh. *Am J Epidemiol* 169:304–312.
12. Frey M, Edwards M (1997) Surveying arsenic occurrence. *J Am Water Works Assoc* 89:105–117.
13. US Environmental Protection Agency (1998) Information available from the safe drinking water information system (US Environmental Protection Agency, Washington, DC), EPA 816-F-98-006.
14. US Environmental Protection Agency (2013) Providing safe drinking water in America: 2013 National Public Water Systems compliance report (US Environmental Protection Agency, Washington, DC), EPA 305-R-15-001.
15. US Census Bureau (2016) Technical documentation: 2016 TIGER/line shapefiles technical documentation (US Census Bureau, Washington, DC).
16. Fritsch F, Carlson R (1980) Monotone piecewise cubic interpolation. *SIAM J Numer Anal* 17:238–246.
17. Leinwand I, Theobald D, Mitchell JKR (2010) Landscape dynamics at the public-private interface: A case study in Colorado. *Landsc Urban Plan* 97:182–193.
18. Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. *Geogr Anal* 24:189–206.
19. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58:267–288.
20. Grooms K (2016) Does water quality improve when a safe drinking water act violation is issued? A study of the effectiveness of the SDWA in California. *BE J Econ Anal Policy* 16:1–23.
21. US Environmental Protection Agency (1999) EPA drinking water and health: What you need to know (US Environmental Protection Agency, Washington, DC), EPA 816-K-99-001.
22. Dziegielewski B, Bik T (2004) Technical assistance needs and research priorities for small community water systems. *J Contemp Water Res Educ* 128:13–20.
23. Tiemann M (2017) *Safe Drinking Water Act (SDWA): A Summary of the Act and Its Major Requirements* (Congressional Research Service, Washington, DC).
24. US Environmental Protection Agency (1999) National water quality inventory (US Environmental Protection Agency, Washington, DC), EPA 841-B-99-005.
25. Bennear L, Olmstead S (2008) The impacts of the "right to know": Information disclosure and the violation of drinking water standards. *J Environ Econ Manage* 56:117–130.
26. Arora S, Cason T (1995) An experiment in voluntary environmental regulation: Participation in EPA's 33/50 program. *J Environ Econ Manage* 28:271–286.
27. Videras J, Alberini A (2000) The appeal of voluntary environmental programs: Which firms participate and why. *Contemp Econ Policy* 18:449–461.
28. Renzetti S, Dupont D (2004) The performance of municipal water utilities: Evidence on the role of ownership. *J Toxicol Environ Health A* 67:1861–1878.
29. Ménard C, Saussier S (2000) Contractual choice and performance the case of water supply in France. *Revue d'économie Industrielle* 92:385–404.
30. Konisky D, Teodoro M (2016) When governments regulate governments. *Am J Pol Sci* 60:559–574.

31. Beecher J, Laubach A (1989) *Survey on Commission Regulation of Water and Sewer Systems* (National Regulatory Research Institute, Columbus, OH).
32. Deen T, et al. (2005) *Ventura and C. Woelfle-Erskine, "Thirsty for Justice: A People's Blueprint for California Water"* (Environmental Justice Coalition for Water, Oakland, CA).
33. National Research Council (1997) Ensuring small water supply system sustainability. *Safe Water from Every Tap: Improving Water Service to Small Communities* (National Academy Press, Washington, DC), p 153.
34. Switzer D, Teodoro M (March 23, 2017) Class, race, ethnicity, and justice in safe drinking water compliance. *Soc Sci Q*, 10.1111/ssqu.12397.
35. National Association of Water Companies (2016) 2016 staff water policy forum (National Association of Water Companies, Washington, DC).
36. US Census Bureau (2011) Decennial census of population, 1940 to 2010 (US Census Bureau, Washington, DC).
37. US Census Bureau (2015) *Population and Housing Unit Estimates* (Population Estimates Program, Washington, DC).
38. US Census Bureau (1992) Population and housing unit counts. Table 2: Population, housing units, area measurements, and density: 1790 to 1990 (US Census Bureau, Washington, DC).
39. US Census Bureau (2013) 2010 census of population and housing: Summary population and housing characteristics (US Census Bureau, Washington, DC), CPH-1-1.
40. US Department of Justice and Federal Trade Commission (2010) *Horizontal Merger Guidelines* (US Department of Justice and Federal Trade Commission, Washington, DC).

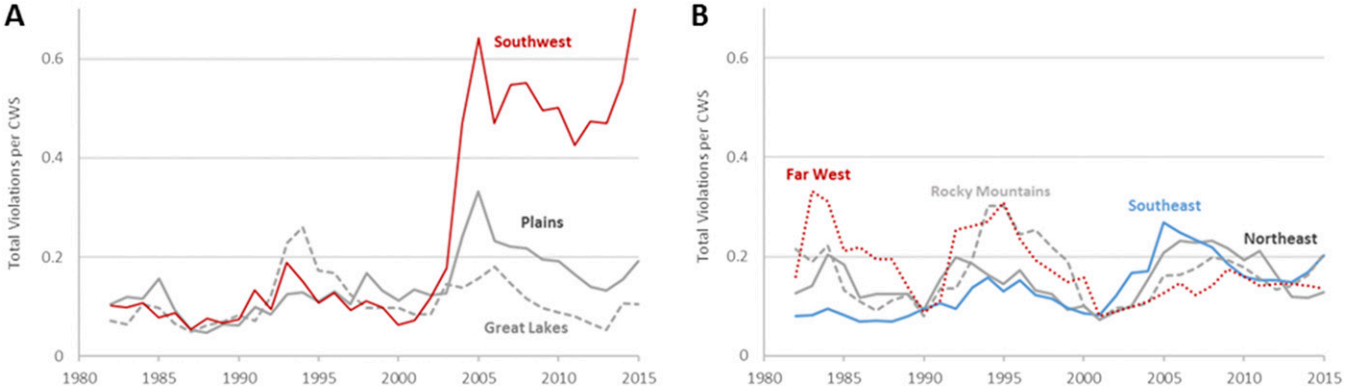

Fig. S1. Health-based violations by region and year. (*A*) Time series for Southwest, Plains, and Great Lakes regions. (*B*) Time series for Far West, Rocky Mountains, Southeast, and Northeast regions. Regions are based on the US Bureau of Economic Analysis (BEA) classifications, as described below. Note: Regions are based on the US BEA classifications. Southwest region includes Arizona, Oklahoma, New Mexico, and Texas. Plains region includes Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota. The Great Lakes includes Illinois, Indiana, Michigan, Ohio, and Wisconsin. Far West comprises California, Nevada, Oregon, and Washington. Rocky Mountains includes Colorado, Idaho, Montana, Utah, and Wyoming. Southeast includes Alabama, Arkansas, Georgia, Florida, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia. The Northeast is a combination of the BEA regions New England and Mideast, which include Connecticut, District of Columbia, Delaware, Massachusetts, Maryland, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont.

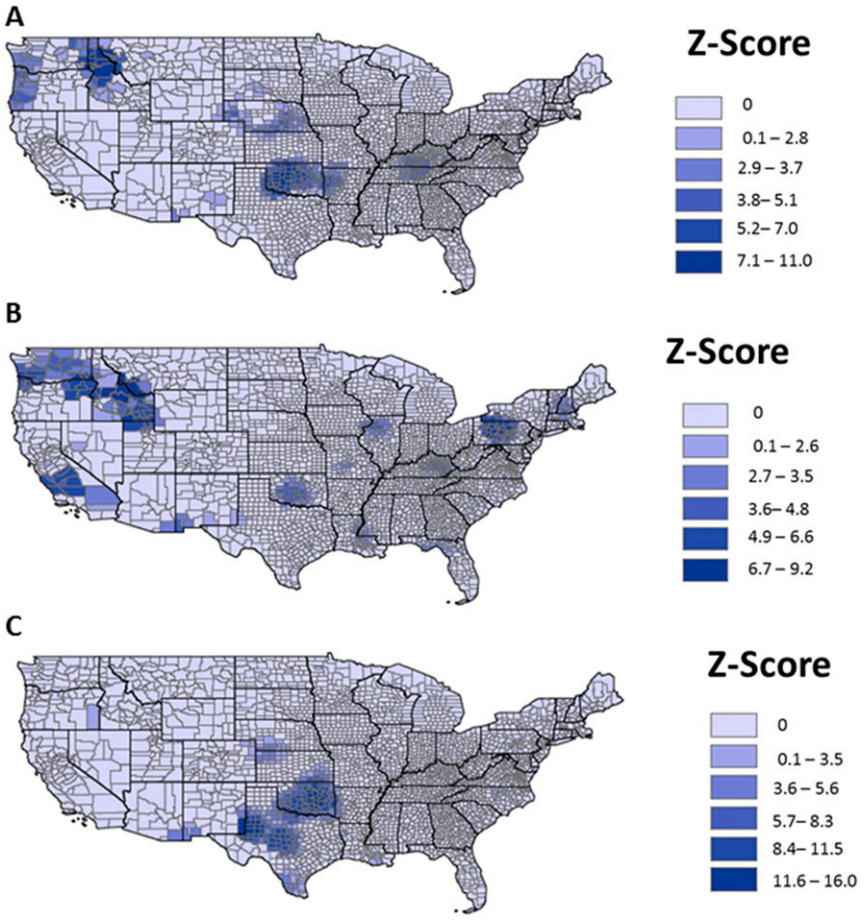

Fig. S2. Spatial clusters (hot spots) of health-based violations, by county and time period. (*A*) Hot-spot analysis for number of total violations per CWS, 1982–1992. (*B*) Hot-spot analysis, 1993–2003. (*C*) Hot-spot analysis, 2004–2015. Intervals in legend are selected based on the Jenks natural breaks classification method.

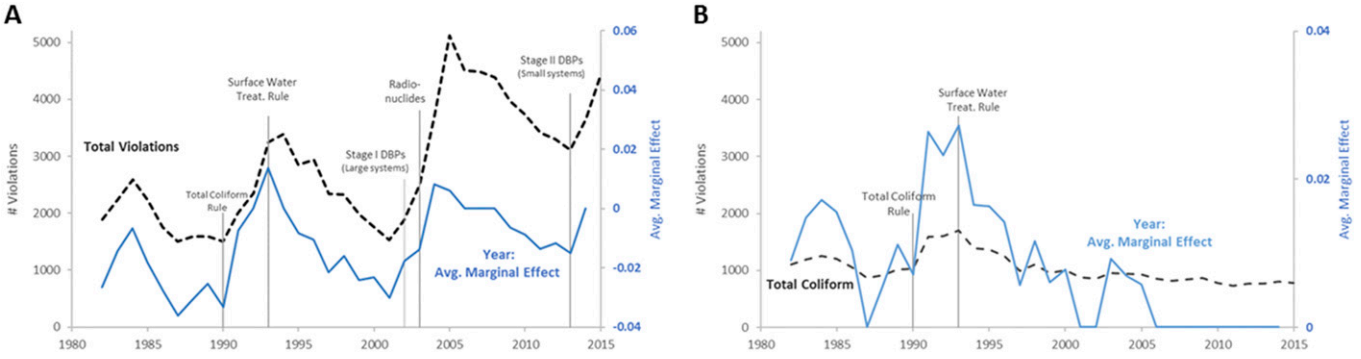

Fig. S3. Average marginal effects (AMEs) of year dummy variables. (*A*) AME for total violations. (*B*) AME for total coliform violations. Note: The secondary *y* axis plots the average marginal effect of year dummy variables. These are estimated in models 2 and 4 for *A* and *B*, respectively.

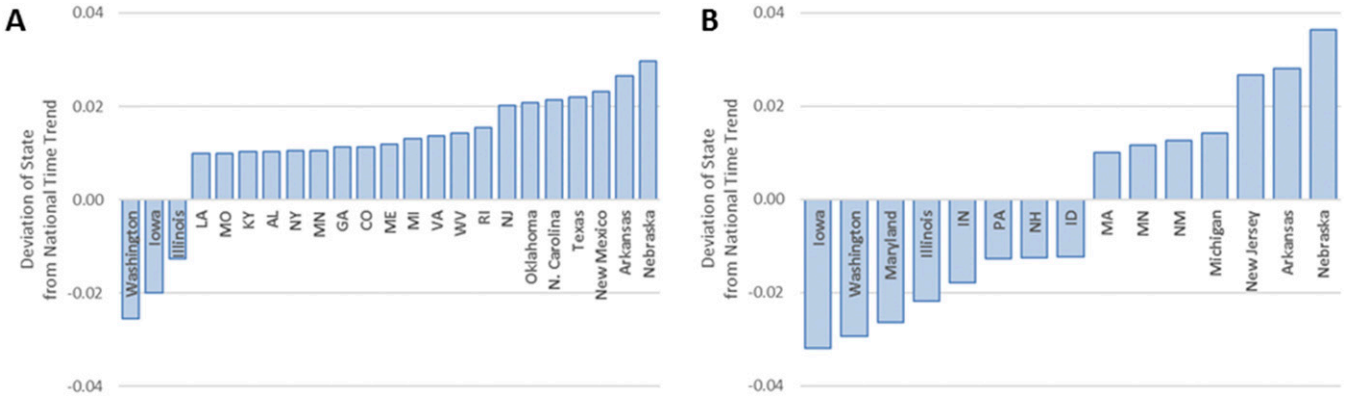

Fig. S4. Deviation of state from national time trend. (*A*) Deviation of state, for total violations. (*B*) Deviation of state, for total coliform violations. Note: Values are the estimated coefficients of the interaction terms (of state dummy variables and the linear time trend). Coefficient values are estimated from models T1 and T3 for *A* and *B*, respectively. Regression results are presented above, in Table S5.

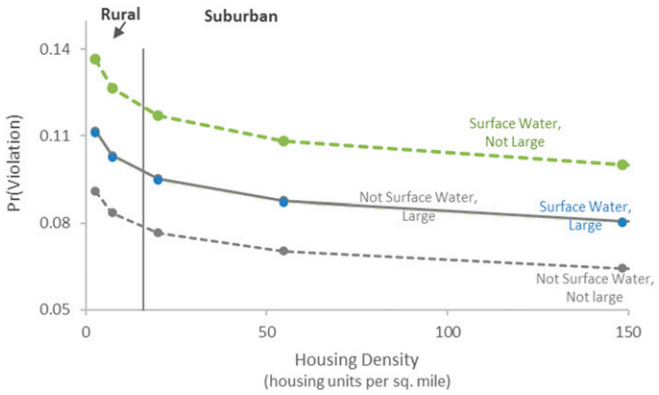

Fig. S5. Average adjusted predictions of water source and system size, across values of housing density.

Table S1. Summary statistics of regression sample

Variable	Description	Mean	SD	Min	Max
Water system-level					
Dependent variables					
All violations	Indicator variable = 1, if water system had any violation in a given year	0.080	0.272	0	1
Total coliform violations	Indicator variable = 1, if water system had total coliform violation in a given year	0.046	0.209	0	1
Independent variables					
Lag violations	Dummy variable = 1, if water system had any violation in previous year	0.080	0.271	0	1
Lag total coliform violations	Dummy variable = 1, if water system had total coliform violation in previous year	0.046	0.210	0	1
Private	Dummy variable = 1, if water system is privately owned	0.140	0.347	0	1
Utility size: small	Dummy variable = 1, if water system serves <3,301 people	0.560	0.496	0	1
Utility size: medium	Dummy variable = 1, if water system serves 3,301–9,999 people	0.227	0.419	0	1
Utility size: large	Dummy variable = 1, if water system serves >10,000 people	0.213	0.409	0	1
Water source: purchased	Dummy variable = 1, if primary water source is purchased water	0.256	0.436	0	1
Water source: surface water	Dummy variable = 1, if primary water source is surface water (not purchased)	0.166	0.372	0	1
Water source: ground water	Dummy variable = 1, if primary water source is ground water (not purchased)	0.579	0.494	0	1
Private*surface water	Interaction: private ownership and surface water source	0.014	0.119	0	1
Private*large	Interaction: private ownership and large-sized utility	0.022	0.146	0	1
Medium*surface water	Interaction: medium-sized utility and surface water source	0.046	0.210	0	1
Large*surface water	Interaction: large-sized utility and surface water source	0.070	0.255	0	1
County level					
Median household income	Median household income, in 2015$	52,211	13,863	13,101	125,624
% Nonwhite	Percent nonwhite population	0.154	0.143	0	0.87
Housing density	Housing units per square mile	146	326	0.2	6,009
HHI (scaled 0–1)	HHI	0.316	0.202	0.03	1
PC1 (“urbanization”)	First PC (for PCA of housing density, income, and % nonwhite)	0.000	1.3	−4.5	4.3
PC2 (“minority, low-income“)	Second PC (for PCA of housing density, income, and % nonwhite)	0.000	1.0	−2.1	5.8
Obs			608,600		
No. CWSs			17,900		

Table S2. Violation occurrence, by CWS characteristics

	All violations		Total coliform		
CWS characteristic	Mean	SD	Mean	SD	Obs
Regression sample	0.080	0.272	0.046	0.209	608,600
Ownership					
Private	0.065	0.247	0.043	0.204	85,102
Local government	0.083	0.275	0.046	0.209	509,728
State/federal/Native American	0.090	0.287	0.061	0.239	13,770
Service population					
Small	0.076	0.265	0.043	0.202	340,714
Medium	0.087	0.282	0.051	0.221	138,346
Large	0.084	0.277	0.049	0.216	129,540
Primary water source					
Purchased water	0.066	0.248	0.036	0.187	155,584
Surface water, not purchased	0.119	0.324	0.033	0.178	100,742
Ground water, not purchased	0.076	0.264	0.054	0.226	352,274
Housing density					
Rural (<16 houses per sq. mi.)	0.095	0.001	0.053	0.001	193,052
Suburban (16–380 houses per sq. mi.)	0.076	0.000	0.044	0.000	364,038
Urban (>380 houses per sq. mi.)	0.054	0.001	0.034	0.001	51,510

Note: Values represent the portion of water system-year observations with violations. For example, for private systems (“Private”), 4.3% of observations have a violation occurrence.

Table S3. Regression results: Coefficient estimates

	All violations						Total coliform					
	(1)			(2)			(3)			(4)		
	No interactions			Interactions			No Interactions			Interactions		
Variable	Coeff.		SE	Coeff.		SE	Coeff.		SE	Coeff.		SE
Lag violation	1.090	***	0.007	1.085	***	0.007	0.795	***	0.009	0.794	***	0.009
Private	−0.077	***	0.008	−0.047	***	0.009	−0.054	***	0.009	−0.035	***	0.010
Medium	0.069	***	0.006	0.080	***	0.007	0.128	***	0.007	0.120	***	0.008
Large	0.061	***	0.007	0.144	***	0.009	0.151	***	0.009	0.178	***	0.010
Water source = purchased	−0.026	***	0.007	−0.034	***	0.007	−0.108	***	0.008	−0.110	***	0.008
Water source = surface water (not purchased)	0.178	***	0.007	0.299	***	0.011	−0.237	***	0.010	−0.224	***	0.017
Private*surface water				−0.075	***	0.023				−0.045		0.033
Private*large				−0.091	***	0.021				−0.086	***	0.024
Medium*surface water				−0.095	***	0.016				0.050	**	0.023
Large*surface water				−0.272	***	0.016				−0.070	***	0.022
ln(median household income)	0.020		0.016	0.017		0.016	−0.014		0.019	−0.015		0.019
% Nonwhite	0.081	***	0.026	0.074	***	0.026	0.186	***	0.030	0.186	***	0.030
ln(housing density)	−0.051	***	0.003	−0.051	***	0.003	−0.034	***	0.003	−0.034	***	0.003
HHI (scaled 0–1)	0.090	***	0.014	0.110	***	0.014	0.058	***	0.017	0.064	***	0.017
Constant	−1.767	***	0.174	−1.763	***	0.175	−1.570	***	0.208	−1.561	***	0.208
Year and state fixed effects	Yes			Yes			Yes			Yes		
Log likelihood	−149,528			−149,361			−103,703			−103,680		
LR χ^2	41,126			41,461			19,379			19,426		
Prob > χ^2	0.000			0.000			0.000			0.000		
McFadden's R^2	0.121			0.122			0.086			0.086		
Obs	608,600			608,600			608,600			608,600		

Note: The general model used to estimate these regressions is presented in the main text, Eq. **1**. ***$P < 0.01$; **$P < 0.05$.

Table S4. Regression results: Marginal effects

	All violations						Total coliform					
	(1)			(2)			(3)			(4)		
	No interactions			Interactions			No interactions			Interactions		
Variable	ME		Delta SE	ME		Delta SE	ME		Delta SE	ME		Delta SE
Lag violation	0.247	***	0.002	0.245	***	0.002	0.120	***	0.002	0.120	***	0.002
Private	−0.010	***	0.001	−0.010	***	0.001	−0.005	***	0.001	−0.005	***	0.001
Medium	0.009	***	0.001	0.008	***	0.001	0.012	***	0.001	0.012	***	0.001
Large	0.008	***	0.001	0.011	***	0.001	0.014	***	0.001	0.015	***	0.001
Water source = purchased	−0.003	***	0.001	−0.004	***	0.001	−0.009	***	0.001	−0.009	***	0.001
Water source = surface water (not purchased)	0.025	***	0.001	0.031	***	0.001	−0.019	***	0.001	−0.018	***	0.001
ln(median household income)	0.003		0.002	0.002		0.002	−0.001		0.002	−0.001		0.002
% Nonwhite	0.011	***	0.003	0.010	***	0.003	0.017	***	0.003	0.017	***	0.003
ln(housing density)	−0.007	***	0.000	−0.007	***	0.000	−0.003	***	0.000	−0.003	***	0.000
HHI (scaled 0–1)	0.012	***	0.002	0.014	***	0.002	0.005	***	0.001	0.006	***	0.001
Year and state fixed effects	Yes			Yes			Yes			Yes		
Log likelihood	−149,528			−149,361			−103,703			−103,680		
LR χ^2	41,126			41,461			19,379			19,426		
Prob > χ^2	0.000			0.000			0.000			0.000		
McFadden's R^2	0.121			0.122			0.086			0.086		
Obs	608,600			608,600			608,600			608,600		

Note: Average marginal effects are not reported for interaction terms since the value of the interaction term cannot change independently of the values of the component terms. ***$P < 0.01$.

Table S5. Time trend regression results: Coefficient estimates

	All violations						Total coliform					
	(T1)			(T2)			(T3)			(T4)		
	Interactions			Interactions			Interactions			Interactions		
Variable	Coeff.		SE	Coeff.		SE	Coeff.		SE	Coeff.		SE
Lag violation	1.085	***	0.007	1.059	***	0.007	0.794	***	0.009	0.763	***	0.009
Private	−0.047	***	0.009	−0.045	***	0.009	−0.035	***	0.010	−0.034	***	0.010
Medium	0.080	***	0.007	0.080	***	0.007	0.120	***	0.008	0.121	***	0.008
Large	0.144	***	0.009	0.145	***	0.009	0.178	***	0.010	0.180	***	0.010
Water source = purchased	−0.034	***	0.007	−0.036	***	0.007	−0.110	***	0.008	−0.112	***	0.008
Water source = surface water (not purchased)	0.299	***	0.011	0.300	***	0.011	−0.224	***	0.017	−0.228	***	0.017
Private*surface water	−0.075	***	0.023	−0.077	***	0.023	−0.045		0.033	−0.045		0.033
Private*large	−0.091	***	0.021	−0.096	***	0.021	−0.086	***	0.024	−0.088	***	0.024
Medium*surface water	−0.095	***	0.016	−0.094	***	0.016	0.050	**	0.023	0.051	**	0.023
Large*surface water	−0.272	***	0.016	−0.273	***	0.016	−0.070	***	0.022	−0.068	***	0.022
ln(median household income)	0.017		0.016	0.000		0.016	−0.015		0.019	−0.032	*	0.019
% Nonwhite	0.074	***	0.026	0.096	***	0.026	0.186	***	0.030	0.189	***	0.030
ln(housing density)	−0.051	***	0.003	−0.051	***	0.003	−0.034	***	0.003	−0.033	***	0.003
HHI (scaled 0–1)	0.110	***	0.014	0.112	***	0.014	0.064	***	0.017	0.065	***	0.017
Year	0.021		0.020	0.013		0.021	−0.002		0.026	−0.002		0.027
Constant	−44.76		39.99	−27.68		41.32	1.660		52.64	3.339		54.28
Year and state fixed effects	Yes			Yes			Yes			Yes		
State-specific time trend	No			Yes			No			Yes		
Log likelihood	−149,361			−148,278			−103,680			−102,734		
LR χ^2	41,461			43,625			19,426			21,318		
Prob > χ^2	0.000			0.000			0.000			0.000		
McFadden's R^2	0.122			0.128			0.086			0.094		
Obs	608,600			608,600			608,600			608,600		

Note: State-specific time trends are interactions of state dummy variables and the linear time trend. The general model to estimate these results is developed based on Eq. **1** in the main text. Two additional terms are added to Eq. **1**—a linear time trend and interactions of the linear time trend and state dummy variables. ***$P < 0.01$; **$P < 0.05$; *$P < 0.10$.

Table S6. PC regression results: Coefficient estimates

	All violations			Total coliform		
	(5)			(6)		
	Interactions			Interactions		
Variable	Coeff.		SE	Coeff.		SE
Lag violation	1.086	***	0.007	0.795	***	0.01
Private	−0.050	***	0.009	−0.037	***	0.01
Medium	0.077	***	0.007	0.118	***	0.01
Large	0.137	***	0.009	0.173	***	0.01
Water source = purchased	−0.043	***	0.007	−0.116	***	0.01
Water source = surface water (not purchased)	0.294	***	0.011	−0.226	***	0.02
Private*surface water	−0.072	***	0.023	−0.043		0.03
Private*large	−0.088	***	0.021	−0.084	***	0.02
Medium*surface water	−0.096	***	0.016	0.049	**	0.02
Large*surface water	−0.274	***	0.016	−0.071	***	0.02
PC1 (“urbanization”)	−0.051	***	0.003	−0.036	***	0.00
PC2 (“minority, low-income”)	−0.006	*	0.003	0.017	***	0.00
HHI (scaled 0–1)	0.113	***	0.014	0.065	***	0.02
Constant	−1.693	***	0.044	−1.780	***	0.05
Year and state fixed effects	Yes			Yes		
Log likelihood	−149,422			−103,702		
LR χ^2	41,339			19,382		
Prob > χ^2	0.000			0.000		
McFadden’s R^2	0.122			0.086		
Obs	608,600			608,600		

Note: Average marginal effects (AMEs) were also estimated for covariates in models 5 and 6. The AME of PC1 is −0.007 (model 5) and −0.003 (model 6). These values are significant at the 1% level. Meanwhile, the AME of PC2 is −0.001 (model 5) and 0.002 (model 6). In model 5, the AME for PC2 is not significant at the 5% level, but in model 6 the AME is significant, at the 1% level. ***$P < 0.01$; **$P < 0.05$; *$P < 0.10$.