Can Collective Action Institutions Outperform the State? Evidence from Treatment of Abandoned Mine Drainage.

Harleman, Max and Weber, Jeremy G.

Georgia College and State University Department of Government and Sociology, University of Pittsburgh Graduate School of Public and International Affairs

22 December 2023

Can Collective Action Institutions Outperform the State? Evidence from Treatment of Abandoned Mine Drainage.*

Max Harleman and Jeremy G. Weber

January 17, 2024

Abstract

A core public administration literature seeks to understand whether decentralized collective action institutions will emerge to provide public goods, such as management of environmental resources. Few studies examine how they perform relative to the state at providing public goods, and they fail to account for the possibility that the state might self-select into providing public goods in the most challenging contexts. If it does, finding that the state performs worse than collective action institutions could reflect its more challenging context rather than differences in knowledge, skill, or motivation. We examine several quantitative measures of performance in remediating polluted water discharges from abandoned coal mines in Pennsylvania, a task sometimes done by the state and sometimes by nonprofit watershed associations. We find that the two types of institutions address discharges with generally similar water quality problems and build systems that yield similar initial improvements in water quality. But watershed association systems better maintain effectiveness at reducing acidity and removing heavy metals over time. The findings suggest a role for sustained public investment in collective action institutions to address complex and enduring environmental problems.

Keywords: collective action, abandoned mines, water quality, decentralization

Word Count: 10,451

*Harleman: Department of Government and Sociology, Georgia College and State University, 410 West Greene Street Campus Box 18. Milledgeville, GA 31061. max.harleman@gcsu.edu. Weber: Graduate School of Public and International Affairs, University of Pittsburgh, 3601 Wesley W. Posvar Hall, Pittsburgh, PA 15260. jgw99@pitt.edu, 3601 Wesley W. Posvar Hall, Pittsburgh, PA 15260. We would like to thank Caroline Fettes and Irving Hu for providing research assistance.

1 Introduction

Private community members have sometimes voluntarily organized into "collective action institutions" to deliver public goods and services, such as the management of common-pool resources like pastures or surface water (North, 1990; Ostrom, 1990). Many scholars and practitioners have argued that collective action institutions provide public goods more efficiently and effectively than public institutions such as federal, state, or local governments. For example, proponents of watershed associations, one type of collective action institution, argue that they are more cost-effective and able to address problems outside the scope of centralized environmental regulation, such as non-point source pollution or habitat destruction (Lubell et al., 2002, p.149). On the other hand, collective action institutions may fail to achieve the success predicted by economic theory (Marshall, 2015; McCord et al., 2017), especially when they attempt to address inherently complex and dynamic policy problems like environmental remediation (North, 1994; Marshall, 2010).

A limited empirical literature concludes that collective action institutions perform better than the state, but it has two notable limitations. The first is a reliance upon case studies (D'Souza and Nagendra, 2011; Mathenge et al., 2014; Wade, 1988; Pahl-Wostl et al., 2012; Cleaver and Toner, 2006; Klassen and Evans, 2020), some of which are chosen because the observed performance of the collective action institutions aligns with their theoretical advantages. Though valuable for their details, such case studies may have limited generalizability even to other collective action institutions working in the same field and region. The second limitation is that, while the literature presents a well-developed theory on why community members organize to provide public goods (Ostrom, 1990, 1999), it does not consider the contextual factors that might lead the state to provide them. For instance, the state might self-select into addressing the most polluted and challenging watersheds, in which case finding that it performs worse than collective action institutions could reflect its more challenging context rather than differences in knowledge, skill, or motivation.

We overcome both limitations in our study of the treatment of abandoned mine drainage in Pennsylvania, a problem that degrades tens of thousands of miles of streams in Appalachian states (Kruse Daniels et al., 2021). The scale of the problem and the two distinct funding sources that address it have produced many treatment systems built by different institutions. The institutional variation stems in large part from a Pennsylvania law that authorizes state grants to non-profit watershed associations that propose systems, and a federal law that provides funds to the state Department of Environmental Protection to implement and oversee treatment. As a result, our sample covers 236 treatment systems, 51 of which were built by the state and the rest by a diversity of watershed associations or related groups.

Our rich data also permit testing for institutional selection, which is our first empirical question: do watershed associations (the collective action institutions) and the state Department of Environmental Protection (the state) specialize in treating discharges of different sizes or severity? Our second empirical question considers cases where the watershed associations and the state address similar discharges and asks whether association systems outperform state systems as measured by monitoring activities, cost-effectiveness, and their initial, average, and long-term effectiveness at cleaning the discharged water.

We find that the state's systems address slightly more acidic discharges. However, many discharges are similar across the two groups of systems. We match each state system with one or more comparable association systems. In the matched sample, the systems treat discharges with similar water quality problems, flow rates, and numbers of discharge points. While the systems perform similarly at cleaning discharge water on average over their observed lifespans, state systems decline in their ability to treat the discharge over time. Conversely, association systems appear to maintain their effectiveness, suggesting that local individuals better recognize that a system is deteriorating and take action to maintain and upgrade systems. We find weaker evidence that state systems are formally monitored more frequently, as measured by the number of laboratory water quality tests taken at the discharge. But informal, on-the-ground monitoring may explain associations' superior maintenance response. In our data, the typical association manages only three geographically concentrated systems, relative to over 50 dispersed systems managed by the state.

Together, our findings illustrate that watershed association systems perform as well as, if not better than, state systems across a variety of outcomes. The finding has contemporary policy relevance. The 2021 Infrastructure Investment and Jobs Act allocates a historic \$11.3 billion for states and tribes to address abandoned mine problems over 15 years, and reauthorizes a federal tax on coal that funds abandoned mine reclamation (Legere, 2021). The increase in funding implies a need for substantial scaling up of reclamation efforts, as it represents a roughly three-fold increase in annual federal funding for abandoned mine problems (US Department of the Interior, 2022). Abandoned mine reclamation is an enduring issue, and systems must be maintained over their long lifespans. Our findings suggest that federal and state agencies can devolve treatment responsibilities to watershed associations, particularly if they also invest in building their organizational capacity. Doing so could help the money go further, as we present suggestive evidence that associations are more cost-effective than the state at treating discharges.

Our findings contribute to two related literatures on public administration. First, by accounting for the state self-selecting into providing public goods in more complex settings and providing four theoretical explanations for such selection (Section 3), we build upon the aforementioned literature on the the relative performance of state and collective action institutions. This literature is exemplified by Ostrom (1965), who studied ground water management in the West Coastal Basin of Southern California in the 1960s. Using a comparative case study approach, Ostrom found that watershed associations were more effective than state-mandated efforts in other basins at enforcing rules to conserve water. Other studies find superior performance of collective action institutions using the cases of agriculture and irrigation systems (Shivakoti et al., 2005; Lam et al., 1998; Wade, 1988), surface water resources (D'Souza and Nagendra, 2011), public water supplies (Cleaver and Toner, 2006; Pahl-Wostl et al., 2012; Mathenge et al., 2014), and management of urban public goods (Foster and Iaione, 2016; Frantzeskaki, 2019).

Second, we expand the empirical literature comparing the performance of public and private organizations across sectors. Public choice theorists have argued that profit-seeking can lead private firms to deliver the quantity and quality of public goods that best meets citizens' demand (Chubb and Moe, 1988). On the other hand, contract failure theorists have argued that public organizations may perform better and be perceived as more trustworthy in providing public goods for which citizens cannot easily assess quantity and quality (Hansmann, 1996). A vast empirical literature explores public versus private performance in the contexts of nursing homes (Broms et al., 2023), hospitals (Rushing, 1974), emergency services (Sobel and Leeson, 2006), and schools (Ballou and Podgursky, 1998; Chubb and Moe, 1988). The studies compare the quality of services before and after they are privatized, or compare state and private organizations that deliver similar goods and services. The approaches have produced mixed findings (see Rainey and Chun (2007)).

A smaller literature finds differences in organizational and managerial characteristics across three sectors: public, private, and nonprofit (Lee and Wilkins, 2011; Koliba et al., 2011; Cohen, 2001; Rainey and Bozeman, 2000; Lan and Rainey, 1992). While the differences could affect performance, Amirkhanyan et al. (2008, p.330) note that most studies of cross-sector performance group public and nonprofit organizations together. We know of only two studies that break out nonprofit and public organizations. Amirkhanyan et al. (2008) find similar performance across public and nonprofit nursing homes in terms of quality of care and accessibility for low-income individuals, and Johansen and Zhu (2013) finds similar levels of performance across public and nonprofit hospitals in terms of their responsiveness to legislative mandates. We expand upon this literature by being the first to compare nonprofit and public performance in the context of environmental public goods, and in this setting find that local collective action institutions may better maintain performance over time.

2 Abandoned Mine Drainage, Funding, and Institutions

2.1 Abandoned Mine Drainage

Since passage of the Federal Surface Mining Control and Reclamation Act of 1977, miners have had to provide financial commitments known as bonds to be allowed to mine. When abandoning a mine, they must comply with reclamation requirements or forfeit their bonds. Prior to the Act, miners could usually abandoned their mines with little to no repercussions. As a result, about half a million abandoned mines with no responsible owner are spread throughout 32 of 50 states in the US (Glatzel and Gordon, 2018), and are unlikely to be addressed apart from public funding.

The problem of abandoned mine hazards is severe in Pennsylvania, which has the largest inventory of unremediated abandoned mines (CRS, 2020), and where the state's Department of Environmental Protection reports that abandoned mine drainage impairs 5,524 miles of stream (Pennsylvania DEP, 2022). The degradation stems from water moving through a mined area and interacting with rock exposed to oxygen through mining. The interaction results in the water often having the same acidity as tomato juice, and containing high concentrations of metals such as iron and aluminum. When it flows into nearby streams, the drainage can kill fish and turn water orange as dissolved iron precipitates. Efforts to address abandoned mine drainage involve a diversity of treatment systems, but they all aim to reduce the acidity of the drainage, capture the metals that precipitate out, and discharge cleaner water into the nearest stream. Most systems are "passive" treatment systems, which typically involve one or more retention ponds or wetlands lined with limestone to reduce acidity and metal contents before the water flows into nearby rivers and streams (US Department of the Interior, 2003). While less expensive than active systems (water treatment plants), they still require large upfront costs to design, construct, and maintain, including costs to replenish limestone or to dredge wetlands that have filled with metals.

2.2 Funding Treatment in Pennsylvania

One major source of funding for treating abandoned mine drainage is set-aside money authorized by the 1990 and 2006 amendments to the federal Surface Mining Reclamation and Control Act. The set-aside provisions allocate money from the federal Abandoned Mine Land Fund, funded by a federal tax on coal, to state or tribal accounts to treat abandoned mine drainage. In the case of Pennsylvania, the Department of Environmental Protection's Bureau of Abandoned Mine Reclamation assumes responsibility for the set-aside money (Pennsylvania DEP, 2023a).[1] This involves

[1] See also Chapter 4-130 of the Federal Assistance Manual of the Office of Surface Mining Reclamation and Enforcement (Office of Surface Mining Reclamation and Enforcement, 2023).

overseeing all phases of a treatment system's life: design, construction, operation, monitoring, and maintenance. The federal Office of Surface Mining Reclamation and Enforcement annually evaluates the Bureau's abandoned mine land program. The Department only expends the set-aside funds in qualified watersheds with a hydrologic plan documenting the health of the watershed, and its topography, human activities, and ecology (Pennsylvania DEP, 2016).

Another major source of funding comes from state watershed protection grants authorized by the Pennsylvania Environmental Stewardship and Watershed Protection Act. The Act directs the Department of Environmental Protection to provide grants to local governments and watershed associations for abandoned mine drainage treatment.[2] The grants have funded work by a variety of watershed associations, which we describe in detail below. Although watershed grants may address restoration of the same streams as the set-aside program, the Bureau of Abandoned Mine Reclamation does not assume legal or maintenance responsibilities for systems built with watershed grants. The Bureau states that it prioritizes use of set-aside funds for operating and maintaining treatment plants or systems "constructed by DEP, or operated by or on behalf of DEP" (Pennsylvania DEP, 2016).

To put a finer point on the implications of the funding streams, with federal set-aside funds the Bureau acts as a funding intermediary and a general contractor. It is accountable to its funder, the federal Office of Surface Mining. With the watershed grants, the Department is the funder, using only state money. It selects which proposals to fund and ensures that recipients comply with its reporting requirements, but it does not act as a general contractor by soliciting subcontractors or putting staff in the field to directly construct or monitor systems. Nor does it report activities and results to a higher authority.

State and association systems are likely built to similar standards because the subcontractors that design and construct systems serve both the state and associations. The Bureau has established guidelines that apply only to its own treatment systems, but they recommend them as a resource for organizations applying for grants (Pennsylvania DEP, 2016). Despite potentially similar design

[2] See section 6105(b)1(iii) of Pennsylvania General Assembly (1999).

and construction quality, performance could vary greatly over time based on whether the associations or the state monitor the systems and address problems as they arise.

2.3 Watershed Associations

Various authors identify a movement in the late 20th century towards managing natural resources at the watershed level (Kenney, 1999; Born and Genskow, 2001). Tarlock (2000) notes that the focus on watershed-level planning came alongside the "emergence of grassroots organizations interested in conserving and restoring specific places" (p.188). This movement was supported at the federal level, with the 1987 amendments to the Clean Water Act authorizing the Environmental Protection Agency to fund community-based efforts to fight non-point source pollution (Hardy and Koontz, 2008).

Watershed associations in Pennsylvania reflect this historical development. Westmoreland County, for example, reports seven main watershed associations, all of which were created between 1970 and 2000 (Westmoreland Conservation District, 2023). Some organizations, like the Babb Creek Watershed Association are staffed by volunteers with no or minimal compensation. Others, like the Mountain Watershed Association, has a full-time professional staff. The associations support diverse activities that promote surface water quality and aquatic life in the watershed such as streambank restoration, fencing, and tree planting. Since abandoned mine drainage affects many of their watersheds, they have also supported treatment efforts, sometimes by pursuing large grants and taking full responsibility for design, construction, and maintenance of treatment systems. A repository of information on treatment systems in Pennsylvania lists roughly 30 different watershed associations or coalitions in its database (Datashed, 2023b). The associations are helped by the Eastern and Western Pennsylvania Coalitions for Abandoned Mine Reclamation, which have full-time, experienced staff that provide technical assistance and serve as a liaison with the Department of Environmental Protection.

3 Theory and Literature

In this section, we use the theoretical and applied literature on collective action to propose four explanations for collective versus state provision of public goods and services. While there is a prolific literature on why non-governmental actors might organize to provide public goods or protect the commons (e.g., Ostrom (1990, 1999); Lubell et al. (2002)), there is comparatively little literature on why the state would provide them without an explicit legislative mandate. We frame our four explanations in the political contracting approach to institutional supply (Libecap, 1989; North, 1990), which suggests that collective action institutions will emerge when their benefits outweigh the transaction costs to organize and maintain them. Our explanations acknowledge that the specific characteristics of the benefits and transaction costs may determine whether the state will provide public goods and crowd out the need for collective action, or otherwise provide them when collective action institutions fail to emerge. We present our explanations as four "personas" of the state, explore their relative ability to provide high-quality public goods versus collective action institutions, and use the explanations to formulate hypotheses about the relative performance of state and association abandoned mine drainage treatment systems. We summarize the four explanations in Table 1.

3.1 The Executor of Last Resort

In our first explanation for the state providing public goods, the government agency is the "Executor of Last Resort." High transaction costs to develop or monitor collective action institutions may preclude their organization. If the state determines that the public good creates benefits that warrant its involvement, it may fill the policy void. This explanation most closely aligns with Salamon (1987), who envisions state action as a response to a failure of collective action, rather than the traditional explanation of state action as a response to market failure (as noted by Amirkhanyan et al. (2008, p.329)). Transaction costs to collective action may be especially high when large numbers of individuals stand to benefit (Olson, 1965). In larger and more diverse groups, individuals have a harder time organizing and agreeing on shared goals, and are incentivized to free-ride.

Transaction costs are also high when little funding is available from the state to support self-governance, such as grants from state and federal governments (Lubell et al., 2002), and in communities with low social and human capital. Communities with low average incomes and low educational attainment typically contain individuals with fewer skills to organize and secure government and private funding, and less disposable time and income to contribute towards organizing (Schneider et al., 1995).

The performance of public goods provided by the Executor of Last Resort (relative to collective action institutions) depends on whether the complexity of providing them is associated with their high transaction costs. For instance, a wealth of environmental justice literature finds that low income communities or communities of color are disproportionately exposed to the most severe environmental harms (e.g., Bullard (1990); Ringquist (2005); Deacon and Baxter (2013); Currie et al. (2023)). An association between community-level variables and problem severity could make it appear that public goods provided by the Executor of Last Resort are less effective or less cost-effective compared to those provided elsewhere by collective action institutions. In terms of on-going monitoring and maintenance of public goods, the comparison between state and collective provision is ambiguous. On one hand, collective action institutions may contain members that live locally and can more easily monitor and provide on-going improvements. On the other, in seeking to fill a known policy void, the state could intentionally enhance their monitoring and maintenance activities.

3.2 The Benevolent Technocrat

In our second explanation, the state is a "Benevolent Technocrat" that chooses to provide public goods when the policy problem is most complex or severe. The Benevolent Technocrat recognizes that public goods are not equally easy to deliver across contexts, and chooses to intervene where the benefits of its involvement is highest. For instance, an environmental regulator may choose to remediate pollution when it is most pervasive or where technical details make cleanup most complex. The idea of the state as a Benevolent Technocrat dates back to the earliest literature in the field of public policy. Lasswell (1956) conceived of a public sector made up of elected officials and teams of policy specialists that would combine political strategy and scientific evidence to design, implement, and evaluate public programs to achieve socially optimal outcomes. Lasswell's conception aligns with later modifications by Simon (1956, 1985), who introduced the concept of "boundedly rational" public officials that "satisfice," or make decisions that they reasonably expect will meet societal goals, even without full knowledge of all possible policy alternatives and their expected outcomes. Perry and Wise (1990) and Perry (2000) similarly view bureaucrats as Benevolent Technocrats, who are predisposed towards fulfilling obligations, maintaining trust, and obeying rules, and are thereby attracted to government service in order to impact socially desirable outcomes.

If the Benevolent Technocrat maintains a staff of qualified professionals, it should have some comparative advantage at delivering public goods relative to local actors. For instance, the state Department of Environmental Protection maintains teams of geologists, hydrologists, construction inspectors, and other experts to execute abandoned mine drainage treatment. But like the Executor of Last Resort, if the Benevolent Technocrat chooses to provide public goods where the policy problem is most severe and complex, it may appear that they are less effective or less cost-effective compared to goods provided elsewhere by collective action institutions. In terms of on-going monitoring and maintenance, the comparison between state and collective provision is again ambiguous.

3.3 The Money Magnet

In our third explanation, the state is a "Money Magnet" that provides public goods because of its comparative advantage at crowding-in funding from many unique sources. Agencies with this persona may accept responsibility for public goods when resources are particularly scarce, or when the cost of providing, monitoring, and maintaining them is particularly high. Under such conditions, transaction costs are high if collective action institutions cannot raise requisite funds. The benefits of goods provision are also high if costliness is associated with the severity of the policy problem. A view of the state as coordinating action across various hierarchical levels of the public, private, and nonprofit sectors aligns with the institutional collective action framework presented by Feiock (2009, 2013) and Feiock et al. (2009). Institutional collective action problems occur when actors dispersed across hierarchies and across sectors pursue similar objectives, as is common in the fields of economic development and environmental management. Government agencies that maintain strong external communication networks may be able to overcome fragmentation by sharing information and expertise, pooling funding sources, and negotiating collaborative agreements that specify who will provide specific components of public goods and services.

If the Money Magnet effectively coordinates across sectors to deliver public goods, their performance may surpass that of collective action institutions in terms of effectiveness, monitoring, and maintenance. If the government agency with primary responsibility for public good provision partners across levels of government and with private and nonprofit entities, it may effectively pool its expertise with local knowledge. In particular, we expect the Money Magnet to exhibit robust monitoring and maintenance due to the combination of technical expertise with local knowledge and presence. However, as in the first two personas, public goods may appear to be less effective and less cost-effective due to self-selection into addressing the most severe and costly policy problems.

3.4 The Obedient Bureaucrat

In our fourth explanation, the state is an "Obedient Bureaucrat." In many policy areas, public goods are too costly for collective institutions to provide without public funds, which may only be accessible under rigid application and reporting requirements that are intended to protect taxpayers. The Obedient Bureaucrat determines that the benefits of public goods are sufficient to warrant its involvement, but its adherence to rigid requirements creates high transaction costs that limit or even intentionally preclude collective action. A view of government as the Obedient Bureaucrat can be traced to Alchian and Demsetz (1972), who point out that rule-based bureaucracies form in the public sector due to lack of market competition that could force managers towards more cost-effective outcomes. While rules help ensure that officials are good stewards of taxpayer dollars, they can also mire public managers in a labyrinth of red tape that replaces the goal of optimizing social welfare with the goal of blindly following rules in order to maintain power, funding, and autonomy (Wilson, 1989). The competing demands of delivering of high quality public goods and minimizing waste are of continued relevance as bureaucracies around the world have aged and undergone reforms (Wise, 2002; Light, 2006).[3]

The performance of the Obedient Bureaucrat relative to collective action institutions is ambiguous with respect to effectiveness and cost-effectiveness, as suggested by our review of the mixed evidence on public versus private action discussed in the introduction. For instance, rules that require contractors to be registered with the state and force the selection of a "lowest responsible bid" may be capable of weeding out overly costly and unprofessional bidders. But they may also unintentionally weed out the most capable or cost-competitive contractors that sort towards private buyers to avoid registering with the state and filling out time-consuming regulatory paperwork. With respect to monitoring and maintenance, we expect that the Obedient Bureaucrat will outperform collective action institutions in cases where these activities are time-consuming and costly. In the case of environmental remediation, monitoring can involve costly field or laboratory tests of air, water, and soil. Maintenance can mean reconstructing existing treatment technologies. State funding decisions often require money to be set aside for such ongoing expenditures that may not be secured upfront in cases of collective provision.

3.5 Hypotheses

Which of the personas best represent the Department of Environmental Protection in their provision of abandoned mine drainage treatment? The answer largely depends on whether it specializes in

[3] For instance, Moynihan (2006) studies New Public Management reforms in US state governments in 1990s, and finds that they were compromised because competing interests led them to be only partially implemented. While reforms related to strategic planning and performance measurement were implemented, almost no states coupled them with the prescribed increases in discretion for public managers related to goal setting, budgeting, hiring, and procurement.

treating the most complex and severe discharges. Our first hypothesis is based off the Executor of Last Resort and Benevolent Technocrat personas, and predicts that the Department will build systems in more complex and severe settings, as measured by the number of contaminated inflows, and the volume and contamination of the water running into its systems compared to the association systems.

Institutional evidence, along with empirical evidence that follows, contradicts the first hypothesis. It suggests that the Department is neither an Executor of Last Resort or a Benevolent Technocrat. The Department constructs many systems in areas with similarly severe water contamination problems and similarly complex technical requirements as those constructed by associations. It also does not build systems to fill a void in areas with low social capital. Instead, it builds systems in the areas with strong watershed associations capable of developing watershed restoration plans using their own resources or state resources earmarked for watershed planning. Moreover, it typically requires associations to provide routine water quality monitoring and maintenance at its systems, including flushing and cleaning retention ponds.

The Department also does not appear to be a pure Money Magnet. While it collaborates with many watershed associations, the empirical evidence that follows shows that it does not choose to construct systems in places that are systematically more expensive. Moreover, although there are many sources of funding for treatment systems, including federal and state programs, private-foundation funding, and in-kind contributions, both state and association systems use a mean of roughly 1.5 funding sources to construct systems (Table 4).

We conclude that the Department is an Obedient Bureaucrat. As discussed in Section 2.2, the federal oversight results in it taking full responsibility for systems built with set-aside funding. Its internal rules reflect the influence of this federal accountability. The Department's Obedient Bureaucratic persona allows us to produce four more hypotheses regarding the performance of its systems versus those built by associations.

Our second hypothesis is that the two types of systems will achieve similar initial and average effectiveness in terms of reducing acidity and removing heavy metals. This hypothesis is motivated by the mixed results from the private versus public performance literature discussed in the introduction, and the institutional details in Section 2.2 that suggest that the two types of systems are likely built to similar standards.

Our third hypothesis is that association systems will be more cost-effective than state systems, as measured by the cost of each gallon of water treated. While contracting requirements may elicit low and very cost-competitive bids, we expect this will be outweighed by greater in-kind contributions of labor by local watershed members, which lower total costs.

Our fourth hypothesis is that state systems are monitored more frequently, as measured by the number of water quality readings taken at the systems over their observed lifespans. Our fifth hypothesis is that state systems will be better maintained, as measured by the change in their effectiveness at improving water quality from their first post-construction reading to their last reading. The Department states that its top priority for the use of federal abandoned mine drainage funding is to operate and maintain its own systems, even above developing new watershed plans and constructing new systems (Pennsylvania DEP, 2016, p.8). It specifically requires local watershed associations to perform routine maintenance (such as site inspections and debris removal), and it intervenes using subsequent allotments of federal funds to carry out non-routine maintenance of failing systems (Pennsylvania DEP, 2016, p.7). Therefore, with its annual stream of federal funds it may outperform the monitoring and maintenance of associations that may run out of money, face challenges in securing grants, or experience deteriorating organization as members move away or age. We note, however, that the number of water quality readings is only one operational definition of monitoring. We do not observe informal monitoring, like dropping by a system to see if effluent appears orange or acidic, which may be more often carried out by local watershed associations.

4 Data and Sample

To test our five hypotheses, we use data on abandoned mine drainage treatment systems from Datashed (2023b). Datashed is an online repository that is funded, operated, maintained by multiple nonprofit and governmental entities (Datashed, 2023a). It contains data on systems in Pennsylvania, including their funding sources, the parties involved in their construction, and a series of water quality variables collected by state employees and association volunteers at multiple points in time. We use variables in Datashed to classify 236 systems as either being state or association constructed and maintained, compare the severity and complexity of the discharges across the two types of systems, and study their differential effects on measures of effectiveness, cost-effectiveness, monitoring, and maintenance.

We note that Datashed provides a system-level report covering all systems and including average influent and effluent water quality readings. This does not permit tracking water quality over time. We therefore download the raw water quality data for each system. From the raw data, we construct influent and effluent water quality readings initially, on average, and at the most recent reading.

4.1 State and Association Systems

We begin with 348 passive systems in Datashed. Of them, 237 have at least one water quality reading to enable our analysis. To classify each system as constructed and maintained by the state or an association, we utilize three variable categories in Datashed. First, each system has a "Project Name" field. We compare the project names to a list of state-constructed passive systems published by Pennsylvania DEP (2019), and classify each system on the list as a state system. Second, each system has fields for its "Responsible Organization" and "Contact Organization." We classify any system where the Department is the responsible organization or contact organization as a state system. Third, each system contains data on the funding sources used to construct the system. We classify any system constructed with federal set-aside money as a state system. We use all three approaches, rather than just one, because some state systems are missing from the Pennsylvania DEP (2019) list, and because there is missingness in the responsible organization, contact organization, and funding fields. We manually checked the individual system web-pages on Datashed, which contain project documents, to verify that each system is properly classified. Altogether, we classify 52 systems as state constructed and maintained and the other 185 as association systems. We drop one state system that is an extreme high outlier for flow volume at 10,955 gallons per minute, with the next highest value being only half this amount, because there is no comparable association system.

4.2 System Characteristics and Water Quality Problems

Table 2 presents descriptive statistics for the systems. The typical system was built between 2003 and 2004 and was funded with 1.4 unique funding sources (i.e., private foundation, federal set-aside, state grants, in-kind contributions). Larger and more complex discharges may be more challenging to treat. To measure complexity, we utilize the number of discharge points treated by the system. The typical system has one discharge, and only systems in the top quartile collect and treat water originating from multiple inflows. To measure size, we use the flow of water in gallons per minute entering the system from all inflows. The average and median systems in our sample take in 278 and 45 gallons of water per minute.

To measure water quality problems, we use lab-conducted water quality tests of water flowing into the systems, known as influent readings. For each system in our sample, we have at least one influent reading that allows us to define severity with a continuous and a threshold-based measure. The continuous measure is the actual value of influent pH, which is a log scale, and concentrations of manganese (Mn), aluminum (Al), total iron (Fe), and total suspended solids (TSS), which are measured in milligrams per liter (mg/L). For pH, which measures acidity, 7 indicates neutral water and lower values indicate greater acidity. Average pH and concentrations of the four metals over all readings taken at the systems are shown in Table 2. For instance, the average influent pH is 4.41, which is calculated by allowing each of the 236 systems in our sample to contribute one pH measure that is its average influent pH taken over all its tests in Datashed. We also calculate an average initial influent pH of 4.48, by taking the average pH from the first reading at each system across the 182 systems that have more than one reading. For these systems, the average number of years between the first and the last reading is 7 years.

Our threshold-based measures are based on regulatory thresholds for pH and the four metal concentrations. The Pennsylvania DEP (2016, p.4) sets water quality targets in streams that receive abandoned mine drainage for pH (greater than 6.0), iron (less than 1.5 mg/l), and aluminum (less than 0.5 mg/l). It lacks targets for the other two measures, so we use targets from the US EPA's coal mining effluent limitation regulations—manganese (less than 2.0 mg/l) and TSS (less than 35 mg/l) (Code of Federal Regulations, 2023). Table 2 shows that 98 percent of the systems treat discharges that fail at least one of the five standards on average over their observed life, with pH (79 percent), iron (77 percent) and aluminum (72 percent) being the most commonly failed standards.

4.3 Effectiveness, Cost-Effectiveness, Monitoring, and Maintenance

Table 3 presents descriptive statistics for the performance outcome variables, which include measures of initial and average effectiveness, cost-effectiveness, monitoring, and long-term maintenance of effectiveness at improving water quality.

As with measures of influent water quality, we use both continuous and threshold approaches to measure system effectiveness. With the continuous measures, we account for initial pollution in influent water by calculating the change in water quality from influent to effluent. For instance, to calculate the change in manganese (Mn) at each system i we calculate $\Delta Ln(Mn_i) = Ln(Mn_{effluent,i}) - Ln(Mn_{influent,i})$. Over the observed life of the average system, effluent had around 66 percent less manganese than influent. The pH scale is logarithmic, so we simply take the change in pH from inflow to outflow and find an average improvement in pH of 1.61 points from influent to effluent. To put the change in perspective, it is the difference between the acidity of tomato juice (4.4) and the acidity of water where aquatic life can flourish.

For the threshold approach to measuring effectiveness, we examine whether the effluent water meets the targets set by the Pennsylvania DEP (2016) for pH, iron, and aluminum, and by the US EPA for manganese and TSS. While 79 percent of systems release effluent that fails any of the five standards, the particular standard that they fail varies. Of the 236 systems in our sample, 36 percent fail the pH standard, between 40 and 50 percent fail the manganese, aluminum, and iron standards, and very few fail the TSS standard.

To measure cost effectiveness, we use the system's initial cost per 100,000 gallons treated. We estimate it by dividing the initial cost of the system by the estimated number of gallons that would run through the system over an assumed 15 year lifespan. We calculate the estimated number of gallons by multiplying the flow rate in gallons per minute by the number of minutes in a 15 year period. On average, systems in our sample cost $241 for each 100,000 gallons of water treated.

To measure monitoring, we use the combined number of water quality readings of both influent and effluent taken over the observed life of the system. Note that effluent readings are those taken from water as it exits the system into the nearest river or stream. The average system has 25 readings taken over its observed lifespan.

To measure maintenance, or the long-term effectiveness of systems at improving water quality, for each system we calculate the most recent measure of effectiveness (effluent less influent at the last reading, $t = T$) and compare it to the initial effectiveness (effluent less influent at the initial reading, $t = 1$). For instance, to calculate the change in manganese at each system i we calculate:

$$\begin{aligned} Change\ \Delta Ln(Mn_{it}) = & (Ln(Mn_{eff,i,t=T}) - Ln(Mn_{inf,i,t=T})) \\ & - (Ln(Mn_{eff,i,t=1}) - Ln(Mn_{inf,i,t=1})) \end{aligned} \tag{1}$$

Table 3 shows that the average system had an initial reduction in manganese that was 0.22 log points higher than the most recent observation. Thus, if the system initially reduced manganese by 40 percent, it is most recently reducing it by around 60 percent, indicating increasing average effectiveness over time. In contrast, the typical system experiences a 16 percent increase in aluminum from inflow to outflow relative to the first reading, indicating decreasing average effectiveness over time. Discharge water quality should generally improve over time as several studies have documented (Burrows et al., 2015; Merritt and Power, 2022). If the state targets discharges from mines abandoned long ago, it should result in its systems treating less acidic or polluted water. This underscores the value of looking at initial inflow water quality across system types and accounting for any differences when estimating treatment effectiveness.

5 Empirical Approach

5.1 Institutional Selection

To answer our first empirical question, which asks whether the state self-selects into addressing the most polluted and challenging discharges, we compare the characteristics of the two types of systems. Specifically, we compare means of the variables in Table 2—those that define system characteristics, complexity, size, and influent severity—using our complete sample of 185 association systems and 51 state systems. We present the actual difference in means and the associated t-statistics for the hypothesis that the two samples come from the same population. We also present the normalized mean difference, which takes a difference in means and divides it by the square root of the average variance of the two groups (Imbens and Wooldridge, 2009). As Imbens and Wooldridge (2009) note, the normalized difference in means is a better measure of covariate differences across two groups when doing causal inference. The traditional t-statistic necessarily increases as the sample size grows, but the normalized difference does not, which is desirable because the challenge of estimating an average treatment effect does not inherently become more difficult as the sample size grows. They note that with a normalized difference in means greater than 0.25 standard deviations, adjusting for the difference through least squares regressions is likely sensitive to model specification.

5.2 Matching and System Performance

The empirical results that follow indicate that in our full sample, state systems treat slightly more acidic discharges. If we were to use our full sample to answer our second empirical question—which asks whether association systems outperform state systems in terms of effectiveness, cost-effectiveness, monitoring, and maintenance—a finding that state systems perform worse could reflect more acidic water being more challenging to treat. Therefore, when testing our second through fifth hypotheses, we address the selection concern in two ways. First, as mentioned in Section 4.3, we define our continuous outcomes as the change in water quality from influent to effluent to account for initial pollution in influent water. Second, we balance the observed characteristics of systems by matching state systems to association systems that address similar discharges.

To match systems, we use a combination of Mahalanobis Distance Matching (MDM) and exact matching. Exact matching matches units across groups, in our case state and association systems, that have the same value for a given variable. MDM matches units across groups that are nearest in covariate space, with nearness defined by a Mahalanobis distance. The Mahalanobis distance is calculated after standardizing each matching variable (which accounts for matching variables being in different units of measurement) and after accounting for correlations between the multiple matching variables (which accounts for redundancy). We use MDM, rather than other popular matching algorithms like Propensity Score Matching (PSM), for two reasons. First, King and Nielsen (2019) show that MDM produces better balance on observed covariates for a given number of observations that are pruned from the full sample. Achieving better balance with a larger sample improves our statistical power to identify differences between state and association systems. Second, Ripollone et al. (2018) find that MDM outperforms PSM when there is a high degree of initial covariate balance by matching observations that are closer in a multidimensional covariate space. Because we have a large sample of association systems that closely overlap our smaller sample of state systems, MDM will yield the closest possible pairs.

The matching procedure can be conceptualized in two phases. First, we begin with the 51 state systems and exact match them based on the variables in Table A1. We exact match on a binary variable indicating whether the water entering the system failed the state standard of 6 pH. This reflects the generally different types of systems needed to treat net acid versus net alkaline mine water (Hedin et al., 2013). The exact match should result in samples of state and association systems with an identical proportion of acidic discharges. We also exact match on a binary variable equal to one if the system is one of the 182 that have more than one water quality reading, which allows us to compare long-term maintenance outcomes across the two types of systems. In the second matching phase, we calculate the Mahalanobis distance for each system based on the eight MDM variables in Table A1. They include system characteristics (year built, the number of inflows, and the inflow volume) and the four continuous measures of influent water quality. Matching on these variables helps balance the size and complexity of the treated discharges and the severity of the influent water quality problem.

Our preferred matching approach is to match each of the 51 state systems to two association systems "without replacement," meaning that the association sample will include 102 unique systems. This approach strikes the best balance between the two opposing goals of reducing observed differences in covariates and retaining a large enough sample to produce statistically significant and externally valid findings on the relationship between institutional choice and performance. In Tables A2 and A3 we present descriptive statistics for all of our covariates and outcome variables for the one-to-two matching without replacement sample. In Appendices B and C we present descriptive statistics and results using a one-to-one without replacement approach and a one-to-two with replacement approach. We discuss the results of the alternative matching methods and why one-to-two matching without replacement is our preferred approach in Section 6.4.

After matching, we compare the means of our outcome variables in Table 3 across the matched samples of state and association systems, again using normalized mean differences. We conclude that the two groups have different population means for our operational definitions of effectiveness, cost-effectiveness, monitoring, and maintenance if their normalized mean differences are greater than roughly one-quarter.

5.3 Matching and Regression-Based Bias Correction

Although we exact match on two variables, the MDM will yield inexact covariate matches for the eight MDM variables. For each of our outcomes in Table 3, we estimate the following model using ordinary least squares, while conditioning the regression on the matched sample:

$$Outcome_i = \beta_1 StateSystem_i + \beta_2 ln(FlowVolume_i) + \beta_3 ln(Inflows_i) + \lambda_i + \varepsilon_{it}, \quad (2)$$

Where i indexes the system, $StateSystem_i$ is a binary variable equal to one if the system is a state system, and $ln(FlowVolume_i)$ and $ln(Inflows_i)$ provide the natural logarithms of two of the MDM variables, influent flow volume in gallons per minute and the number of inflow points. We also control for construction year fixed-effects, λ_i. With the year fixed-effects, our coefficient of interest β_1 estimates the average difference in $Outcome_i$ across state and association systems built in the same year.[4] We estimate the equation with Huber-White heteroscedasticity robust standard errors. For the monitoring and cost-effectiveness outcomes, we estimate equation 2 as written. For the effectiveness and binary effectiveness threshold outcomes, we estimate parameters in the equation by weighting each observations using analytical weights of the number of readings that contribute to its average pH and pollutant concentrations.[5] For the maintenance outcomes, we additionally control for $ln(MeasurmentDuration_i)$, which is the numbers of years between the first and the last reading taken at system i.

Note that we control for the first three of our eight MDM variables in equation 2. For the other five, we directly account for differences in influent water quality by defining our continuous outcomes as the change in water quality from influent to effluent. For the non-continuous, binary outcomes, we condition each regression on failure of the influent water quality standard. For instance, for the binary indicator of effluent manganese being greater than 2 mg/l, we only include observations in the regression if the influent water is greater than 2 mg/l. This avoids the inclusion of observations in regressions for a given outcome where the effect of the system is nonbinding. We do not estimate the equation for the TSS outcome because too few systems have influent that fails the standard to provide a sample large enough for the estimation of the model parameters.

[4]We use the "reghdfe" command in Stata to estimate equation 2. It does not report a constant, β_0. We therefore do not report a constant in the proceeding tables, which would simply be the mean outcome of the omitted build year.

[5]The exception is with the one-to-two matching with replacement sample in Appendix C. With this sample, we use frequency weights to weight each state system by one, and the 65 association systems by the number of times they are matched to a state system.

6 Results

6.1 Institutional Selection

We begin by testing our first hypothesis that the state will build systems in more complex and severe settings, using our full sample of state and association systems. In the full sample of 185 association systems and 51 state systems, the association systems are about three years newer on average (Table 4). The two types of systems have a similar number of funding sources. They also address similarly complex and similarly sized bodies of water, as measured by the number of inflows and inflow volume. The clearest difference across the two types of systems are for measures of influent acidity. Both systems have an average influent acidity of between 4 and 5 in their first reading and on average over their observed lifespan. But 92 percent of state systems treat discharges with an influent pH of less than the state standard of 6 compared to only 76 percent of association systems, a normalized mean difference of .46 and nearly two times the one-quarter threshold.

Moving to our matched samples using one-to-two matching without replacement, association systems are still two years newer than state systems on average, but the difference is smaller in the matched sample relative to the full sample (Table 5). Most notably, exact matching on a binary for failing the influent pH standard along with MDM on influent pH yields samples with similar influent acidity.

Together, Tables 4 and 5 provide empirical evidence to test our first hypothesis, which predicts that the state will build systems in the most complex and severe settings, as measured by the number and volume of the inflows and the severity of influent water quality. The two types of entities treat similar discharges in nearly every dimension, except for slight differences in influent acidity (Table 4). But the state systems are very similar on all dimensions, including the influent acidity, to a matched sample of 102 association systems (55 percent of all the association systems in our full sample) (Table 5). The weight of the empirical evidence contradicts our first hypothesis, because the state constructs systems with similar characteristics, similarly severe water quality problems, and similarly complex technical requirements as many of the systems constructed by watershed associations.

That the Department of Environmental Protection constructs many systems that treat similar discharges suggests that it is neither an Executor of Last Resort or a Benevolent Technocrat that aims to fill a void in the most challenging institutional settings. Along with the institutional evidence in Section 2, rejecting our first hypothesis lends support for the state as an Obedient Bureaucratic, with its internal rules for watershed planning directing where it builds systems, rather than the severity of the contamination issue.

6.2 Matching and System Performance

To examine the relative performance of state and association systems, we use mean comparisons across the state and association systems for each of our operational definitions of effectiveness, cost-effectiveness, monitoring, and maintenance. State and association systems have relatively similar levels of average effectiveness over their observed lifespans (Table 6). The exception is with removing iron from the influent water, for which state systems are slightly more effective by about a third of a standard deviation. While average performance at increasing pH is similar over the observed lifespans of the systems, state systems are slightly more effective at initially increasing pH above the state threshold at the first reading, also by about a third of a standard deviation.

State and association systems appear to differ in both cost-effectiveness and monitoring activity. For cost-effectiveness, state systems' cost to treat 100,000 gallons of water is 37 percent higher on average, although the difference is statistically imprecise. For monitoring activity, the state conducts a significantly greater number of water quality readings, a difference that is better identified by the proceeding biased-corrected regressions that account for differences in the amount of time for readings to occur by conditioning on the year fixed-effects.

The largest difference between state and association systems is the maintenance of long-term effectiveness. The means for the state systems consistently indicate declining effectiveness over time. The negative sign on the change in pH effectiveness indicates declining ability to increase pH over time, and the positive signs on the change in the natural logarithm of manganese, aluminum, and iron indicate increasing concentrations of metals over time.[6] Association systems, on the other hand, appear to be maintaining or even improving their performance at increasing pH and removing metals over time.

6.3 Matching and Regression-Based Bias Correction

While our matched samples are similar on most observed covariates, an exception is with the year built variable (Table 5). To address the concern that we are comparing newer association systems to older state systems, we estimate regressions including year fixed-effects that effectively compare state and association systems built in the same year. We also control for inexact covariate matches on the complexity and size controls, and condition the effectiveness threshold regressions on failing the influent water quality standard for a given effluent outcome.

For the effectiveness outcomes, the two types of systems show similar performance at meeting state standards, raising pH, and removing metals from influent (Table 7). Taken together, the results for effectiveness in Tables 6 and 7 support our second hypothesis, which states that the two types of systems will achieve similar initial and average effectiveness in terms of increasing pH and removing heavy metals.

State systems are 37 percent more costly than association systems, but the difference is imprecise (Table 8). Because the magnitude of the difference is large, we cannot rule out our third hypothesis which predicted that association systems would be more cost-effective than state systems. This may indicate that for some systems there are inefficiencies from state contracting requirements (or efficiencies from in-kind contributions of labor by local watershed members), but that these differences are not ubiquitous.

State systems receive 36 percent more water quality readings than association systems, but

[6] Of the five maintenance outcome means for state systems, two are statistically different from zero at the 95 percent level using a two-tailed t-test: the means for the change in pH and the change in aluminum. The means for association systems are not statistically different from zero using a two-tailed test.

this difference is also imprecise. We therefore cannot rule out our fourth hypothesis that state systems would be monitored more frequently due to their explicit preference for expending annual allocations of set-aside funds to monitor and maintain their systems, but there are at least some association systems that achieve similar levels of monitoring.

For the maintenance outcomes, association systems continue to show stronger performance relative to state systems when accounting for the bias correction (Table 9). The negative coefficient on the “state” binary in the pH regression indicates that association systems are superior to state systems at maintaining acidity reductions by nearly 1 pH scale point. The positive signs on the “state” binary in the other four regressions present imprecise evidence that state systems are declining in effectiveness at removing metals and suspended solids from effluent over time, relative to the same change at association systems.

Appendix Figures A1 through A3 display overlapping histograms across state and association systems for the three maintenance outcomes with the strongest effects. They visually display that the distribution of association systems are experiencing superior performance at raising pH and removing aluminum and manganese over time, relative to the state systems.

Together, the results from the three figures, the mean comparisons (Table 6) and the bias-corrected regressions (Table 9) contradict our fifth hypothesis that state systems will be better maintained. This is despite the Department of Environmental Protection’s prioritization of the use federal funds to operate and maintain its existing systems (Pennsylvania DEP, 2016, p.8). Instead, it appears that the primary advantage of decentralized institutions described in the collective action literature, namely the immediate presence of local individuals to informally monitor the quality of environmental resources and intervene when necessary, results in superior maintenance of water quality in streams and rivers near the discharges over time.

6.4 Robustness

In Appendices B and C we present descriptive statistics and results using a one-to-one without replacement approach and a one-to-two with replacement matching approach. These approaches achieve very modest improvements in covariate balance, but with smaller sample sizes of 51 and 65 unique association systems this comes at the expense of statistical precision and external validity. Implicit in our second empirical question is that we are concerned with the relative performance of some large set of association systems and state systems—rather than simply producing the most internally valid estimate of one of the two treatment conditions. Because one-to-two matching without replacement produces the largest sample of unique systems that still balances our covariates, we believe it is the matching approach that strikes the best balance between internal and external validity.

The results in Appendices B and C are substantively similar to the one-to-two without replacement approach presented here in the main text. Where there are differences, the biased-corrected regressions provide more precise evidence that state systems are more effective at bringing effluent above state aluminum standards (Tables B6 and C6), association systems are more cost-effective (Table B6), and association systems outperform state systems at maintaining their effectiveness at removing aluminum (Table C7).

7 Discussion and Policy Implications

Our main finding is that associations better maintain their systems' effectiveness over time. For insight into why, we qualitatively examine associations that best maintained effectiveness. The Babb Creek Watershed Association in Tioga County is responsible for four of the best fifteen systems with above 95th percentile improvements in pH, as well as aluminum and manganese concentrations. Babb Creek built its twelve systems in our dataset in the late 1990s and early 2000s. They treat highly acidic discharges, with eleven out of twelve below the state standard of 6 pH at the first influent reading. As of the most recent effluent readings, ten of twelve are bringing pH above the state standard. Clearfield Creek Watershed Association has also successfully maintained its systems. Its two systems in our dataset that treat discharges into Little Laurel Run in Cambria County had an average initial influent pH of 3.5, and the most recent effluent readings are well above the state standard of 6 pH. Similarly, Blacklick Creek Watershed Association manages seven systems in Cambria and Indiana counties, with an average initial influent pH of 4. As of the most recent readings, all seven were bringing effluent pH above the state standard.

Although the three associations are entirely run by volunteers, they bring in between $100,000 and $600,000 in revenues in the typical year through private and public contributions (ProPublica, 2022b,c,a). While some of the contributions are small, like the sale of raffle tickets and the proceeds from festivals, fishing tournaments, and community events, others can be quite large. The three associations' financial reports show occasional sharp increases in revenues and expenditures in years where they receive large private donations or grants to build new systems or maintain their current systems. For instance, Babb Creek Watershed Association received a grant of $186,000 in 2019 through the state's Environmental Stewardship and Watershed Protection Act to maintain one of its systems built in 2004 (Pennsylvania DEP, 2023b). Successful fundraising and maintenance activities of the three associations point to the traditional theoretical advantages of collective action institutions (e.g., (Ostrom, 2005)), with local individuals volunteering their time to manage natural resources, building social capital and community around the natural resource, and informally monitoring environmental quality and intervening when necessary.

Another advantage lies in the ability to fund-raise quickly or apply for emergency funds, which may be a quicker method for repairing failing systems relative to rigid state procurement rules. For example, through the Quick Response Emergency Repair Funding program, which is funded by the Department of Environmental Protection and a private energy company, the Western Pennsylvania Coalition for Abandoned Mine Reclamation provides associations with money to restore failing systems. The Coalition states that in the most urgent cases, the money can arrive in a few days (Western Pennsylvania Coalitions for Abandoned Mine Reclamation, 2023). All three of the associations discussed in this section have received Quick Response funds, which helps explain their resiliency.

Despite having many well-maintained systems, the state is responsible for some systems that significantly declined in effectiveness over time (see Figure A1). Its four systems with the worst performance at maintaining pH improvements were built between 1998 and 2000. Two of the four have no record of ongoing maintenance in Datashed, and one has not been significantly modified since 2010. The fourth, which was built and initially effective at raising pH in 1999, was only recently maintained and updated in 2022, well after the most recent water quality reading in 2018. The state's seemingly slow response may be because of red tape surrounding procurement processes, or because they have over 50 systems statewide compared to the smaller, geographically concentrated number managed by each association. It may also be due to declining money in the federal Abandoned Mine Land Fund, which comes from the federal excise tax on coal production. The US Department of the Interior (2023) reports that coal excise tax revenue has fallen nearly every year since the start of the shale oil and gas boom that precipitated a decline in domestic coal production. Between 2008 and 2022, the decline in coal production reduced excise tax revenues by 73 percent (US Energy Information Administration, 2023; US Department of the Interior, 2023).

More generally, relying on collective action institutions to address environmental problems in the US may be important in light of stagnant funding for state environmental protection agencies, which are increasingly tasked with doing more with less (Cusick, 2017). The agencies are asked to implement and enforce federal and state programs to protect air, water, and land that are constantly in flux due to changing political administrations. But available funds to run programs and pay staff, like administrators, inspectors, and technical experts, appear to be decreasing. For example, the Pennsylvania Department of Environmental Protection had a total allocation of $786 million from all funding sources in 2019 (Commonwealth of Pennsylvania, 2021, p.E18-6), which is about 30 percent less in real terms than the Department's historically high allocations at the turn of the century (Commonwealth of Pennsylvania, 2002, p.E16.7).

As states and tribes prepare to spend the historic $11.3 billion earmarked in the 2021 Infrastructure Investment and Jobs Act for abandoned mine reclamation, our results suggest that the state should continue to rely on and invest in its collective action institutions to scale up reclamation efforts. While in the past states like Pennsylvania have assumed full responsibility for spending federal funds to address abandoned mine issues (Pennsylvania DEP, 2023a), we find that watershed associations are at least as effective at addressing mine discharges. The many systems that will be built with Infrastructure Act money must be maintained over their lifespans, which may be 25 years or more. The state can leverage watershed associations by investing in their organizational capacity and devolving more responsibility to construct and maintain systems. Doing so could help the money go further. We present suggestive evidence that associations are more cost-effective than the state, and the actual difference might be even larger than we estimate. While the state project costs include direct labor costs, such time spent by contractors building the systems, they exclude overhead labor costs associated with administration. By comparison, watershed associations rely in part on local volunteers. To the extent that people enjoy volunteering to improve water quality in their community, some of their labor represents a local benefit rather than a cost.

8 Conclusion

Our findings illustrate that sustained public investment in collective action institutions can help address complex and enduring environmental problems left by legacy hazards. State systems to treat mine drainage treat slightly more polluted water than do systems managed by watershed associations. Many discharges, however, are similar across the two types of systems. Comparing systems that treat similar discharges, we find association systems perform at least as well as state systems. They have similar performance at cleaning discharge water initially and on average over their observed lifespans. Over time, state systems decline on average in their ability to treat the discharge, while association systems better maintain their effectiveness. Suggestive evidence also indicates that association systems are more cost-effective. As such, current federal appropriations for addressing abandoned mine drainage and other hazards might go further—and maybe materialize more quickly—through increased partnerships with watershed associations or similar nonprofit partners.

References

Alchian, A. A. and H. Demsetz (1972). Production, Information Costs, and Economic Organization. *The American Economic Review 62*(5), 777–795.

Amirkhanyan, A. A., H. J. Kim, and K. T. Lambright (2008). Does the public sector outperform the nonprofit and for-profit sectors? Evidence from a national panel study on nursing home quality and access. *Journal of Policy Analysis and Management 27*(2), 326–353.

Ballou, D. and M. Podgursky (1998). Teacher recruitment and retention in public and private schools. *Journal of Policy Analysis and Management 17*(3), 393–417.

Born, S. M. and K. D. Genskow (2001). Toward understanding new watershed initiatives. *University of Wisconsin-Madison (en internet, http://clean-water. uwex. edu/initiatives/watershed. pdf)*.

Broms, R., C. Dahlström, and M. Nistotskaya (2023, 03). Provider Ownership and Indicators of Service Quality: Evidence from Swedish Residential Care Homes. *Journal of Public Administration Research and Theory*, muad002.

Bullard, R. D. (1990). *Dumping in Dixie: Race, class, and environmental quality*. Avalon Publishing-(Westview Press).

Burrows, J. E., S. C. Peters, and C. A. Cravotta (2015). Temporal geochemical variations in above- and below-drainage coal mine discharge. *Applied Geochemistry 62*, 84–95. A Special Issue to Honor the Geochemical Contributions of D. Kirk Nordstrom.

Chubb, J. E. and T. M. Moe (1988). Politics, Markets, and the Organization of Schools. *American Political Science Review 82*(4), 1065–1087.

Cleaver, F. and A. Toner (2006). The evolution of community water governance in Uchira, Tanzania: The implications for equality of access, sustainability and effectiveness. *Natural Resources Forum 30*(3), 207–218.

Code of Federal Regulations (2023). 40 CFR 434.22: Effluent limitation guidelines representing the degree of effluent reduction attainable by the application of the best practicable control technology currently available (BPT). `https://www.ecfr.gov/current/title-40/chapter-I/subchapter-N/part-434`.

Cohen, S. (2001). A Strategic Framework for Devolving Responsibility and Functions from Government to the Private Sector. *Public Administration Review 61*(4), 432–440.

Commonwealth of Pennsylvania (2002). Governor's Executive Budget 2002-2003. `https://www.budget.pa.gov/Publications%20and%20Reports/Documents/2002_03_Budget.pdf`.

Commonwealth of Pennsylvania (2021). Governor's Executive Buget 2021-2022. `https://www.budget.pa.gov/Publications%20and%20Reports/CommonwealthBudget/Documents/2021-22%20Proposed%20Budget/2021-2022%20Executive%20Budget%20Book.Web%20Version.UPDATED.030421.pdf`.

CRS (2020). (Congressional Research Service) The Abandoned Mine Reclamation Fund: Reauthorization Issues in the 116th Congress. `https://crsreports.congress.gov/product/pdf/R/R46266`.

Currie, J., J. Voorheis, and R. Walker (2023, January). What Caused Racial Disparities in Particulate Exposure to Fall? New Evidence from the Clean Air Act and Satellite-Based Measures of Air Quality. *American Economic Review 113*(1), 71–97.

Cusick, M. (2017). Cash-Strapped State Environmental Agencies Brace For Budget Cuts. *National Public Radio*.

Datashed (2023a). About Datashed. `https://www.datashed.org/help/3478110`.

Datashed (2023b). Datashed Projects. `https://www.datashed.org/`.

Deacon, L. and J. Baxter (2013). No opportunity to say no: a case study of procedural environmental injustice in Canada. *Journal of Environmental Planning and Management 56*(5), 607–623.

D'Souza, R. and H. Nagendra (2011). Changes in public commons as a consequence of urbanization: The Agara lake in Bangalore, India. *Environmental management 47*, 840–850.

Feiock, R. C. (2009). Metropolitan Governance and Institutional Collective Action. *Urban Affairs Review 44*(3), 356–377.

Feiock, R. C. (2013). The Institutional Collective Action Framework. *Policy Studies Journal 41*(3), 397–425.

Feiock, R. C., A. Steinacker, and H. J. Park (2009). Institutional Collective Action and Economic Development Joint Ventures. *Public Administration Review 69*(2), 256–270.

Foster, S. R. and C. Iaione (2016). The City as a Commons. *Yale Law and Policy Review 34*(2), 281–349.

Frantzeskaki, N. (2019). Seven lessons for planning nature-based solutions in cities. *Environmental Science & Policy 93*, 101–111.

Glatzel, M. and B. Gordon (2018). The West's Sleeping Giant: Abandoned Mines and the role of the Good Samaritan. `https://waterinthewest.stanford.edu/news-events/news-insights/wests-sleeping-giant-abandoned-mines-and-role-good-samaritan`.

Hansmann, H. (1996). The changing roles of public, private, and nonprofit enterprise in education, health care, and other human services. In *Individual and social responsibility: Child care, education, medical care, and long-term care in America*, pp. 245–276. University of Chicago Press.

Hardy, S. D. and T. M. Koontz (2008). Reducing nonpoint source pollution through collaboration: policies and programs across the US States. *Environmental Management 41*, 301–310.

Hedin, R., T. Weaver, N. Wolfe, and G. Watzlaf (2013). Effective passive treatment of coal mine drainage. In *Proceedings of the 35th National association of abandoned mine land programs conference, Daniels, WV, USA*.

Imbens, G. W. and J. M. Wooldridge (2009, March). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature 47*(1), 5–86.

Johansen, M. and L. Zhu (2013, 06). Market Competition, Political Constraint, and Managerial Practice in Public, Nonprofit, and Private American Hospitals. *Journal of Public Administration Research and Theory 24*(1), 159–184.

Kenney, D. S. (1999). Historical and Sociolpolitcal Context of the Western Watersheds Movement. *JAWRA Journal of the American Water Resources Association 35*(3), 493–503.

King, G. and R. Nielsen (2019). Why propensity scores should not be used for matching. *Political analysis 27*(4), 435–454.

Klassen, S. and D. Evans (2020). Top-down and bottom-up water management: A diachronic model of changing water management strategies at Angkor, Cambodia. *Journal of Anthropological Archaeology 58*, 101166.

Koliba, C. J., R. M. Mills, and A. Zia (2011). Accountability in Governance Networks: An Assessment of Public, Private, and Nonprofit Emergency Management Practices Following Hurricane Katrina. *Public Administration Review 71*(2), 210–220.

Kruse Daniels, N., J. A. LaBar, and L. M. McDonald (2021). Acid mine drainage in Appalachia: sources, legacy, and treatment. *Appalachia's Coal-Mined Landscapes: Resources and Communities in a New Energy Era*, 193–216.

Lam, W. F. et al. (1998). *Governing irrigation systems in Nepal: institutions, infrastructure, and collective action.* Institute for Contemporary Studies.

Lan, Z. and H. G. Rainey (1992, 01). Goals, Rules, and Effectiveness in Public, Private, and Hybrid Organizations: More Evidence on Frequent Assertions About Differences. *Journal of Public Administration Research and Theory 2*(1), 5–28.

Lasswell, H. D. (1956). The Decision Process: Seven Categories of Functional Analysis. *Bureau of Government Research*.

Lee, Y.-j. and V. M. Wilkins (2011). More Similarities or More Differences? Comparing Public and Nonprofit Managers' Job Motivations. *Public Administration Review 71*(1), 45–56.

Legere, L. (2021). Infrastrcture bill promises 'unprecdented'funding to clean up Pa.'s abandoned mine lands. *Pittsburgh Post-Gazette. November 11*.

Libecap, G. D. (1989). *Contracting for Property Rights*. Cambridge University Press.

Light, P. C. (2006). The Tides of Reform Revisited: Patterns in Making Government Work, 1945–2002. *Public Administration Review 66*(1), 6–19.

Lubell, M., M. Schneider, J. T. Scholz, and M. Mete (2002). Watershed Partnerships and the Emergence of Collective Action Institutions. *American Journal of Political Science 46*(1), 148–163.

Marshall, G. R. (2010). Governance for a surprising world. *Resilience and transformation: Preparing Australia for uncertain futures*, 49–57.

Marshall, G. R. (2015). Polycentricity, subsidiarity and adaptive efficiency. In *Workshop on Polycentricity*. Ostrom Workshop, Indiana University.

Mathenge, J., C. N. Luwesi, C. A. Shisanya, I. Mahiri, R. A. Akombo, and M. N. Mutiso (2014). Community participation in water sector governance in Kenya: A performance based appraisal of community water management systems in Ngaciuma-Kinyaritha catchment, Tana basin, Mount Kenya region. *International Journal of Innovative Research and Development 3*(5), 783–92.

McCord, P., J. Dell'Angelo, E. Baldwin, and T. Evans (2017). Polycentric Transformation in Kenyan Water Governance: A Dynamic Analysis of Institutional and Social-Ecological Change. *Policy Studies Journal 45*(4), 633–658.

Merritt, P. and C. Power (2022). Assessing the long-term evolution of mine water quality in abandoned underground mine workings using first-flush based models. *Science of The Total Environment 846*, 157390.

Moynihan, D. P. (2006). Managing for Results in State Government: Evaluating a Decade of Reform. *Public Administration Review 66*(1), 77–89.

North, D. C. (1990). *Institutions, Institutional Change, and Economic Performance.* Cambridge University Press.

North, D. C. (1994). Economic Performance Through Time. *The American Economic Review 84*(3), 359–368.

Office of Surface Mining Reclamation and Enforcement (2023). Acid Mine Drainage Set-Aside Program. `https://www.osmre.gov/sites/default/files/pdfs/4-130.pdf`.

Olson, M. (1965). The Logic of Collective Action. *Contemporary Sociological Theory 124*.

Ostrom, E. (1965). *Public entrepreneurship: a case study in ground water basin management*. University of California, Los Angeles.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Ostrom, E. (1999). Institutional rational choice: an assessment of the institutional analysis and development framework theories of the policy process. *Theories of the Policy Process. Boulder, Colarado: Westview*, 35–72.

Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.

Pahl-Wostl, C., L. Lebel, C. Knieper, and E. Nikitina (2012). From applying panaceas to mastering complexity: Toward adaptive water governance in river basins. *Environmental Science & Policy 23*, 24–34.

Pennsylvania DEP (2016). (Department of Environmental Protection) Acid Mine Drainage Set-Aside Program Program Implementation Guidelines. `http://www.depgreenport.state.pa.us/elibrary/GetDocument?docId=8133&DocName=ACID%20MINE%20DRAINAGE%20SET-ASIDE%20PROGRAM%20IMPLEMENTATION%20GUIDELINES.PDF%20`.

Pennsylvania DEP (2019). Passive Abandoned Mine Drainage (AMD) Treatment Systems Constructed by PA-DEP-Bureau of Abandoned Mine Reclamation.

Pennsylvania DEP (2022). (Department of Environmental Protection) Integrated Water Quality Report – 2022. `https://www.dep.pa.gov/Business/Water/CleanWater/WaterQuality/IntegratedWatersReport/Pages/2022-Integrated-Water-Quality-Report.aspx`.

Pennsylvania DEP (2023a). (Department of Environmental Protection) Acid Mine Drainage (AMD) Program.

Pennsylvania DEP (2023b). Investment Tracker. `http://www.dced.state.pa.us/InvestmentTracker/DefaultDEP.aspx`.

Pennsylvania General Assembly (1999). Environmental Stewardship and Watershed Protection Act. `https://www.legis.state.pa.us/cfdocs/legis/LI/consCheck.cfm?txtType=HTM&ttl=27&div=0&chpt=61`.

Perry, J. L. (2000). Bringing society in: Toward a theory of public-service motivation. *Journal of public administration research and theory 10*(2), 471–488.

Perry, J. L. and L. R. Wise (1990). The motivational bases of public service. *Public administration review*, 367–373.

ProPublica (2022a). Babb Creek Watershed Association Inc. `https://projects.propublica.org/nonprofits/organizations/232995774`.

ProPublica (2022b). Blacklick Creek Watershed Association Inc. `https://projects.propublica.org/nonprofits/organizations/251763928`.

ProPublica (2022c). Clearfield Creek Watershed Association. `https://projects.propublica.org/nonprofits/organizations/251883781`.

Rainey, H. G. and B. Bozeman (2000, 04). Comparing Public and Private Organizations: Empirical Research and the Power of the A Priori. *Journal of Public Administration Research and Theory 10*(2), 447–470.

Rainey, H. G. and Y. H. Chun (2007, 06). 71 Public and Private Management Compared. In *The Oxford Handbook of Public Management*. Oxford University Press.

Ringquist, E. J. (2005). Assessing evidence of environmental inequities: A meta-analysis. *Journal of Policy Analysis and Management 24*(2), 223–247.

Ripollone, J. E., K. F. Huybrechts, K. J. Rothman, R. E. Ferguson, and J. M. Franklin (2018, 05). Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. *American Journal of Epidemiology 187*(9), 1951–1961.

Rushing, W. (1974). Differences in Profit and Nonprofit Organizations: A Study of Effectiveness and Efficiency in General Short-Stay Hospitals. *Administrative Science Quarterly 19*(4), 474–484.

Salamon, L. M. (1987). *Partners in Public service: The scope and theory of government nonprofit relations. In W. W. Powell (Ed.), The nonprofit sector: A research handbook*. New Haven, CT: Yale University Press.

Schneider, M., P. Teske, and M. Mintrom (1995). *Public Entrepreneurs: Agents for Change in American Government*. Princeton University Press.

Shivakoti, G., D. Vermillion, W.-F. Lam, E. Ostrom, U. Pradhan, and R. Yoder (2005). *Asian irrigation in transition: Responding to challenges*. SAGE Publishing India.

Simon, H. A. (1956). Rationality as Process and as Product of Thought. *The American Economic Review*.

Simon, H. A. (1985). Human Nature in Politics: The Dialogue of Psychology with Political Science. *American Political Science Review*.

Sobel, R. S. and P. T. Leeson (2006). Government's response to Hurricane Katrina: A public choice analysis. *Public Choice 127*, 55–73.

Tarlock, D. (2000). Putting rivers back in the landscape: The revival of watershed management in the United States. *UC Law Environmental Journal 6*(2), 167.

US Department of the Interior (2003). Technical Note 409: Passive Treatment Systems for Acid Mine Drainage. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1018&context=usblmpub#:~:text=PASSIVE%20TREATMENT%20SYSTEMS%20provide%20a,Department%20of%20Environmental%20Protection%201999).

US Department of the Interior (2022). Interior Department Announces $122.5 Million for Abandoned Mine Land Economic Revitalization Grants. www.doi.gov/pressreleases/interior-department-announces-1225-million-abandoned-mine-land-economic-revitalization.

US Department of the Interior (2023). Coal Excise Tax. https://revenuedata.doi.gov/how-revenue-works/coal-excise-tax/.

US Energy Information Administration (2023). Coal Data Broswer: Aggregate Coal Mine Production. https://www.eia.gov/coal/data/browser/.

Wade, R. (1988). *Village Republics: Economic Conditions for Collective Action in South India*. Cambridge: Cambridge University Press.

Western Pennsylvania Coalitions for Abandoned Mine Reclamation (2023). Quick Response for Funding Emergency Repairs. http://www.wpcamr.org/projects/QuickResponse/index.html.

Westmoreland Conservation District (2023). Watershed Associations. http://westmorelandconservation.org/watershed-associations/.

Wilson, J. Q. (1989). *Bureaucracy: What Government Agencies Do And Why They Do It*. New York: Basic Books.

Wise, L. R. (2002). Public Management Reform: Competing Drivers of Change. *Public Administration Review 62*(5), 556–567.

Table 1: Four Personas of the State

	The Executor of Last Resort	The Benevolent Technocrat	The Money Magnet	The Obedient Bureaucrat
Transaction Costs to Collective Action	High (preclude organization)	Low to High	High (difficult to raise requisite funds)	High (stringent rules)
Benefits to Public Goods	Medium to High	High (most complex cases)	High (most expensive and complex)	Medium to High
Implications for Performance				
Effectiveness	May appear worse if transaction costs are associated with problem severity	Better, but may appear worse due to complexity	Better, but may appear worse due to complexity	May be better or worse
Cost-Effectiveness	May appear worse if transaction costs are associated with costs of provision	May appear worse due to complexity	Likely better, but may appear worse due to complexity	May be better or worse
Monitoring	May be better or worse	May be better or worse	Likely better	Likely better
Maintenance	May be better or worse	May be better or worse	Likely better	Likely better

Table 2: Treatment System Descriptive Statistics

	Mean	SD	Min.	p25	p50	p75	Max	N
Characteristics	.	.	.	.	.	.	.	.
State System	0.22	0.41	0.00	0.00	0.00	0.00	1.00	236.00
Year Built	2,003.42	6.45	1,970.00	1,999.00	2,004.00	2,007.00	2,019.00	236.00
Number of Funding Sources	1.38	0.69	1.00	1.00	1.00	2.00	6.00	236.00
Severity	.	.	.	.	.	.	.	.
Initial Influent pH	4.48	1.42	2.50	3.30	3.88	6.10	7.60	182.00
Avg Influent pH	4.41	1.39	2.49	3.19	3.85	5.87	7.79	236.00
Influent Manganese (mg/l)	10.35	18.38	0.00	1.53	3.00	9.76	119.33	236.00
Influent Aluminum (mg/l)	12.21	34.49	0.00	0.40	3.46	12.56	465.21	236.00
Influent Iron (mg/l)	38.03	108.23	0.00	1.73	9.85	39.60	1,512.67	236.00
Influent TSS (mg/l)	24.97	130.03	0.00	5.00	7.59	16.94	1,944.50	236.00
Severity Threshold	.	.	.	.	.	.	.	.
Any Influent Standard Failed	0.98	0.13	0.00	1.00	1.00	1.00	1.00	236.00
Initial Influent pH ≤ 6	0.73	0.45	0.00	0.00	1.00	1.00	1.00	182.00
Avg Influent pH ≤ 6	0.79	0.41	0.00	1.00	1.00	1.00	1.00	236.00
Influent Mang. ≥ 2 (mg/l)	0.65	0.48	0.00	0.00	1.00	1.00	1.00	236.00
Influent Aluminum $\geq .5$ (mg/l)	0.72	0.45	0.00	0.00	1.00	1.00	1.00	236.00
Influent Iron ≥ 1.5 (mg/l)	0.77	0.42	0.00	1.00	1.00	1.00	1.00	236.00
Influent TSS ≥ 35 (mg/l)	0.10	0.30	0.00	0.00	0.00	0.00	1.00	236.00
Complexity	.	.	.	.	.	.	.	.
Number of Inflows	1.26	0.58	1.00	1.00	1.00	1.00	4.00	236.00
Size	.	.	.	.	.	.	.	.
Inflow Volume (gal/min)	277.60	774.33	0.53	18.36	44.97	161.25	7,239.38	236.00

Note: Data are from Datashed water quality reports and project documents.

Table 3: Treatment System Descriptive Statistics, Outcome Variables

	Mean	SD	Min.	p25	p50	p75	Max	N
Effectiveness	.	.	.	.	.	.	.	.
Initial Change in pH	1.96	1.68	-2.76	0.60	2.07	3.41	4.95	182.00
Avg Change in pH	1.61	1.58	-2.68	0.35	1.15	3.06	4.72	236.00
Δ Ln Manganese (mg/l)	-0.66	1.08	-7.04	-1.03	-0.46	-0.11	2.79	233.00
Δ Ln Aluminum (mg/l)	-1.48	1.72	-6.86	-2.47	-1.23	-0.11	4.45	231.00
Δ Ln Iron (mg/l)	-1.75	1.89	-6.83	-2.84	-1.56	-0.50	6.48	234.00
Δ Ln TSS (mg/l)	-0.10	0.99	-4.45	-0.61	0.00	0.49	2.36	228.00
Effectiveness Threshold	.	.	.	.	.	.	.	.
Any Effluent Standard Failed	0.79	0.41	0.00	1.00	1.00	1.00	1.00	236.00
Initial Effluent pH ≤ 6	0.21	0.41	0.00	0.00	0.00	0.00	1.00	182.00
Avg Effluent pH ≤ 6	0.36	0.48	0.00	0.00	0.00	1.00	1.00	236.00
Effluent Mang. ≥ 2 (mg/l)	0.48	0.50	0.00	0.00	0.00	1.00	1.00	236.00
Effluent Aluminum $\geq .5$ (mg/l)	0.45	0.50	0.00	0.00	0.00	1.00	1.00	236.00
Effluent Iron ≥ 1.5 (mg/l)	0.42	0.50	0.00	0.00	0.00	1.00	1.00	236.00
Effluent TSS ≥ 35 (mg/l)	0.05	0.21	0.00	0.00	0.00	0.00	1.00	236.00
Cost Effectiveness	.	.	.	.	.	.	.	.
Cost Per 100k Gallons Treated	241.10	895.72	0.00	19.83	70.87	186.77	10,477.85	224.00
Monitoring	.	.	.	.	.	.	.	.
Number of Readings	25.30	39.61	2.00	6.00	11.50	32.00	402.00	236.00
Maintenance	.	.	.	.	.	.	.	.
Change in pH Effectiveness	-0.05	1.32	-4.46	-0.62	-0.03	0.40	4.03	182.00
Change in Δ Ln Manganese	-0.22	1.37	-5.96	-0.73	-0.06	0.40	5.00	165.00
Change in Δ Ln Aluminum	0.15	1.53	-4.01	-0.70	0.00	0.99	6.11	160.00
Change in Δ Ln Iron	0.05	1.48	-5.18	-0.69	0.08	0.93	4.37	161.00
Change in Δ Ln TSS	-0.07	1.33	-4.25	-0.81	0.00	0.66	3.98	145.00

Note: Data are from Datashed water quality reports and project documents.

Table 4: Comparison of State and Association Systems

	Association System (N=185)	State System (N=51)	Diff.	Normalized Diff.	P-Value
Characteristics	.	.	.	.	.
Year Built	2,004.02	2,001.25	2.77	0.50	0.00
Number of Funding Sources	1.34	1.53	-0.19	-0.26	0.12
Severity	.	.	.	.	.
Initial Influent pH	4.61	4.00	0.61	0.45	0.01
Avg Influent pH	4.49	4.13	0.36	0.27	0.09
Influent Manganese (mg/l)	10.97	8.10	2.88	0.18	0.20
Influent Aluminum (mg/l)	11.46	14.92	-3.45	-0.11	0.46
Influent Iron (mg/l)	39.24	33.65	5.58	0.06	0.62
Influent TSS (mg/l)	28.26	13.08	15.18	0.15	0.17
Severity Threshold	.	.	.	.	.
Any Influent Standard Failed	0.98	1.00	-0.02	-0.21	0.05
Influent pH ≤ 6	0.69	0.87	-0.19	-0.46	0.01
Avg Influent pH ≤ 6	0.76	0.92	-0.16	-0.46	0.00
Influent Mang. ≥ 2 (mg/l)	0.65	0.67	-0.02	-0.04	0.81
Influent Aluminum $\geq .5$ (mg/l)	0.72	0.73	-0.00	-0.00	0.99
Influent Iron ≥ 1.5 (mg/l)	0.77	0.78	-0.02	-0.04	0.80
Influent TSS ≥ 35 (mg/l)	0.11	0.06	0.05	0.18	0.23
Complexity	.	.	.	.	.
Number of Inflows	1.25	1.29	-0.04	-0.07	0.69
Size	.	.	.	.	.
Inflow Volume (gal/min)	277.39	278.34	-0.94	-0.00	0.99

Note: Normalized mean differences are calculated using the method of Imbens and Wooldridge (2009), which is the difference in means divided by the square root of the average variance. P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Severity Threshold variables are all binary variables.

Table 5: Comparison of One-to-Two Without Replacement MDM Matched State and Association Systems

	Association System (N=102)	State System (N=51)	Diff.	Normalized Diff.	P-Value
Characteristics	.	.	.	.	.
Year Built	2,003.40	2,001.25	2.15	0.48	0.00
Number of Funding Sources	1.41	1.53	-0.12	-0.16	0.35
Severity	.	.	.	.	.
Initial Influent pH	4.06	4.00	0.06	0.05	0.81
Avg Influent pH	4.06	4.13	-0.07	-0.06	0.75
Influent Manganese (mg/l)	6.77	8.10	-1.33	-0.12	0.51
Influent Aluminum (mg/l)	10.68	14.92	-4.23	-0.20	0.29
Influent Iron (mg/l)	23.68	33.65	-9.98	-0.22	0.21
Influent TSS (mg/l)	12.68	13.08	-0.40	-0.02	0.91
Severity Threshold	.	.	.	.	.
Any Influent Standard Failed	0.98	1.00	-0.02	-0.20	0.16
Initial Influent pH ≤ 6	0.85	0.87	-0.03	-0.07	0.71
Avg Influent pH ≤ 6	0.92	0.92	0.00	0.00	1.00
Influent Mang. ≥ 2 (mg/l)	0.65	0.67	-0.02	-0.04	0.81
Influent Aluminum $\geq .5$ (mg/l)	0.84	0.73	0.12	0.29	0.11
Influent Iron ≥ 1.5 (mg/l)	0.70	0.78	-0.09	-0.20	0.24
Influent TSS ≥ 35 (mg/l)	0.07	0.06	0.01	0.04	0.81
Complexity	.	.	.	.	.
Number of Inflows	1.23	1.29	-0.07	-0.11	0.52
Size	.	.	.	.	.
Inflow Volume (gal/min)	219.61	278.34	-58.73	-0.09	0.58

Note: Normalized mean differences are calculated using the method of Imbens and Wooldridge (2009), which is the difference in means divided by the square root of the average variance. P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Severity Threshold variables are all binary variables.

Table 6: Mean Differences: One-to-Two Without Replacement MDM Matched State and Association Systems

	Association System (N=102)	State System (N=51)	Diff.	P-Value
Effectiveness	.	.	.	.
Initial Change in pH	2.05	2.71	-0.66	0.03
Avg Change in pH	1.84	1.71	0.13	0.63
Δ Ln Manganese (mg/l)	-0.62	-0.66	0.04	0.84
Δ Ln Aluminum (mg/l)	-1.72	-1.46	-0.26	0.35
Δ Ln Iron (mg/l)	-1.51	-2.12	0.60	0.05
Δ Ln TSS (mg/l)	0.04	0.11	-0.06	0.70
Effectiveness Threshold	.	.	.	.
Any Effluent Standard Failed	0.75	0.80	-0.05	0.49
Initial Effluent pH ≤ 6	0.29	0.15	0.14	0.07
Avg Effluent pH ≤ 6	0.41	0.47	-0.06	0.50
Effluent Mang. ≥ 2 (mg/l)	0.46	0.49	-0.03	0.73
Effluent Aluminum $\geq .5$ (mg/l)	0.46	0.59	-0.13	0.14
Effluent Iron ≥ 1.5 (mg/l)	0.36	0.39	-0.03	0.73
Effluent TSS ≥ 35 (mg/l)	0.05	0.04	0.01	0.78
Cost Effectiveness	.	.	.	.
Ln Cost 100k Gallons Treated	3.98	4.34	-0.37	0.19
Monitoring	.	.	.	.
Ln Number of Readings	2.48	2.93	-0.46	0.03
Maintenance	.	.	.	.
Change in pH Effectiveness	0.21	-0.52	0.73	0.01
Change in Δ Ln Manganese	-0.40	0.28	-0.68	0.02
Change in Δ Ln Aluminum	0.01	0.89	-0.89	0.01
Change in Δ Ln Iron	-0.09	0.20	-0.29	0.49
Change in Δ Ln TSS	-0.17	-0.02	-0.16	0.64

Note: P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Effectiveness Threshold variables are all binary variables.

Table 7: Biased Corrected Effects: Effectiveness, One-to-Two Without Replacement

	Any	Initial pH		pH		Mn		Al		Fe	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	p(Fail)	Δ	p(Fail)	Δ	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)
State System	0.06	0.66	-0.18	-0.33	0.11	0.14	-0.08	0.31	0.20	-0.03	0.16
	(0.08)	(0.35)	(0.12)	(0.32)	(0.12)	(0.21)	(0.13)	(0.32)	(0.10)	(0.35)	(0.13)
Ln Influent Flow (gal/min)	0.03	0.03	-0.03	-0.06	0.01	0.10	0.07	0.13	0.01	0.20*	0.02
	(0.02)	(0.10)	(0.03)	(0.08)	(0.03)	(0.07)	(0.05)	(0.08)	(0.03)	(0.09)	(0.04)
Ln Number of Inflows	-0.11	0.10	-0.04	0.41	-0.27*	-0.72**	0.01	-1.01*	-0.33*	-0.85	0.06
	(0.12)	(0.43)	(0.14)	(0.41)	(0.13)	(0.26)	(0.13)	(0.44)	(0.14)	(0.47)	(0.12)
R-squared	0.11	0.20	0.10	0.16	0.12	0.22	0.18	0.22	0.24	0.19	0.27
N	144	110	92	146	134	144	93	141	117	144	104
Cond. on Failed Influent Std.	Y	N	Y	N	Y	N	Y	N	Y	N	Y

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. Each system in the regressions is weighted using analytical weights of the number of readings that contribute its average pH and pollutant concentrations.

Table 8: Biased Corrected Effects: Cost-Effectiveness and Monitoring, One-to-Two Without Replacement

	Ln(Cost Per 100k Gallons)	Ln(Number of Readings)
	(1)	(2)
State System	0.37	0.36
	(0.32)	(0.22)
Ln Number of Inflows	0.92*	0.61*
	(0.40)	(0.26)
Ln Influent Flow (gal/min)		0.05
		(0.07)
R-squared	0.20	0.26
N	143	146

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. We do not control for influent flow in gallons per minute in column 1, because it is directly used to estimate the demoninator of the cost per 100,000 gallons outcome.

Table 9: Biased Corrected Effects: Maintenance, One-to-Two Without Replacement

	Change in Δ pH	Change in Δ Ln			
	(1)	(2)	(3)	(4)	(5)
	pH	Mn	Al	Fe	TSS
State System	-0.99**	0.54	0.71	0.14	0.36
	(0.36)	(0.39)	(0.37)	(0.53)	(0.47)
Ln Influent Flow (gal/min)	-0.03	-0.03	0.10	0.05	-0.07
	(0.09)	(0.09)	(0.12)	(0.14)	(0.13)
Ln Number of Inflows	-0.36	0.67*	0.07	0.52	0.00
	(0.32)	(0.32)	(0.46)	(0.48)	(0.40)
Ln Measurement Duration	-0.43*	0.15	0.05	0.07	0.01
	(0.21)	(0.30)	(0.23)	(0.30)	(0.30)
R-squared	0.21	0.17	0.25	0.13	0.19
N	110	96	91	95	82

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction.

Appendices For Online Publication

A Additional Tables and Figures

Table A1: Matching Variables

Variable	Description
Exact Matching	
Influent pH ≤ 6	Mean influent pH is less than state standard of 6
Missingness of time varying variables	Binary variable indicating one of the 182 systems with more than one sample
Mahalanobis Distance Matching (MDM)	
Year Built	Year the system was built
Number of Inflows	Number of inflows into the system
Inflow Volume	Total inflow into the system in gallons per minute
Influent pH	Mean influent pH on the pH scale
Influent Manganese (mg/l)	Mean influent Manganese in mg/l
Influent Aluminum (mg/l)	Mean influent Aluminum in mg/l
Influent Iron (mg/l)	Mean influent Iron in mg/l
Influent TSS (mg/l)	Mean influent Suspended Solids in mg/l

Table A2: Descriptive Statistics: One-to-Two Without Replacement MDM Sample

	Mean	SD	Min.	p25	p50	p75	Max	N
Characteristics	.	.	.	.	.	.	.	.
State System	0.33	0.47	0.00	0.00	0.00	1.00	1.00	153.00
Year Built	2,002.69	4.74	1,988.00	1,999.00	2,003.00	2,005.00	2,016.00	153.00
Number of Funding Sources	1.45	0.69	1.00	1.00	1.00	2.00	4.00	153.00
Severity	.	.	.	.	.	.	.	.
Initial Influent pH	4.04	1.21	2.50	3.10	3.65	4.46	7.21	117.00
Avg Influent pH	4.08	1.19	2.49	3.14	3.65	4.93	6.76	153.00
Influent Manganese (mg/l)	7.21	11.29	0.00	1.53	2.97	7.79	64.75	153.00
Influent Aluminum (mg/l)	12.09	18.88	0.00	0.80	6.02	13.79	144.62	153.00
Influent Iron (mg/l)	27.00	43.60	0.00	1.04	7.67	31.85	228.50	153.00
Influent TSS (mg/l)	12.81	19.31	0.00	4.70	6.50	12.83	152.00	153.00
Severity Threshold	.	.	.	.	.	.	.	.
Any Influent Standard Failed	0.99	0.11	0.00	1.00	1.00	1.00	1.00	153.00
Initial Influent pH ≤ 6	0.85	0.35	0.00	1.00	1.00	1.00	1.00	117.00
Avg Influent pH ≤ 6	0.92	0.27	0.00	1.00	1.00	1.00	1.00	153.00
Influent Mang. ≥ 2 (mg/l)	0.65	0.48	0.00	0.00	1.00	1.00	1.00	153.00
Influent Aluminum $\geq .5$ (mg/l)	0.80	0.40	0.00	1.00	1.00	1.00	1.00	153.00
Influent Iron ≥ 1.5 (mg/l)	0.73	0.45	0.00	0.00	1.00	1.00	1.00	153.00
Influent TSS ≥ 35 (mg/l)	0.07	0.25	0.00	0.00	0.00	0.00	1.00	153.00
Complexity	.	.	.	.	.	.	.	.
Number of Inflows	1.25	0.59	1.00	1.00	1.00	1.00	4.00	153.00
Size	.	.	.	.	.	.	.	.
Inflow Volume (gal/min)	239.18	620.37	0.53	18.89	49.67	153.15	5,637.25	153.00

Note: Data are from Datashed water quality reports and project documents.

Table A3: Descriptive Statistics: One-to-Two Without Replacement MDM Sample, Outcome Variables

	Mean	SD	Min.	p25	p50	p75	Max	N
Effectiveness	.	.	.	.	.	.	.	.
Initial Change in pH	2.27	1.62	-2.10	0.80	2.70	3.60	4.95	117.00
Avg Change in pH	1.80	1.54	-1.97	0.56	1.40	3.19	4.72	153.00
Δ Ln Manganese (mg/l)	-0.63	1.12	-7.04	-0.96	-0.44	-0.12	2.79	151.00
Δ Ln Aluminum (mg/l)	-1.63	1.60	-6.86	-2.71	-1.34	-0.41	1.37	148.00
Δ Ln Iron (mg/l)	-1.71	1.80	-6.42	-2.84	-1.49	-0.49	3.86	151.00
Δ Ln TSS (mg/l)	0.06	0.88	-2.55	-0.44	0.07	0.65	2.36	146.00
Effectiveness Threshold	.	.	.	.	.	.	.	.
Any Effluent Standard Failed	0.77	0.42	0.00	1.00	1.00	1.00	1.00	153.00
Initial Effluent pH ≤ 6	0.25	0.43	0.00	0.00	0.00	0.00	1.00	117.00
Avg Effluent pH ≤ 6	0.43	0.50	0.00	0.00	0.00	1.00	1.00	153.00
Effluent Mang. ≥ 2 (mg/l)	0.47	0.50	0.00	0.00	0.00	1.00	1.00	153.00
Effluent Aluminum $\geq .5$ (mg/l)	0.50	0.50	0.00	0.00	1.00	1.00	1.00	153.00
Effluent Iron ≥ 1.5 (mg/l)	0.37	0.49	0.00	0.00	0.00	1.00	1.00	153.00
Effluent TSS ≥ 35 (mg/l)	0.05	0.21	0.00	0.00	0.00	0.00	1.00	153.00
Cost Effectiveness	.	.	.	.	.	.	.	.
Cost Per 100k Gallons Treated	222.09	689.78	0.15	20.85	76.96	175.87	7,268.57	150.00
Monitoring	.	.	.	.	.	.	.	.
Number of Readings	28.12	45.73	2.00	6.00	14.00	36.00	402.00	153.00
Maintenance	.	.	.	.	.	.	.	.
Change in pH Effectiveness	-0.03	1.51	-4.46	-0.74	-0.08	0.50	4.03	117.00
Change in Δ Ln Manganese	-0.20	1.31	-3.81	-0.78	-0.13	0.42	5.00	103.00
Change in Δ Ln Aluminum	0.28	1.67	-4.01	-0.66	0.18	1.16	6.11	100.00
Change in Δ Ln Iron	0.00	1.73	-5.18	-0.92	0.17	1.09	4.37	102.00
Change in Δ Ln TSS	-0.13	1.35	-3.08	-1.04	-0.04	0.69	3.29	90.00

Note: Data are from Datashed water quality reports and project documents.

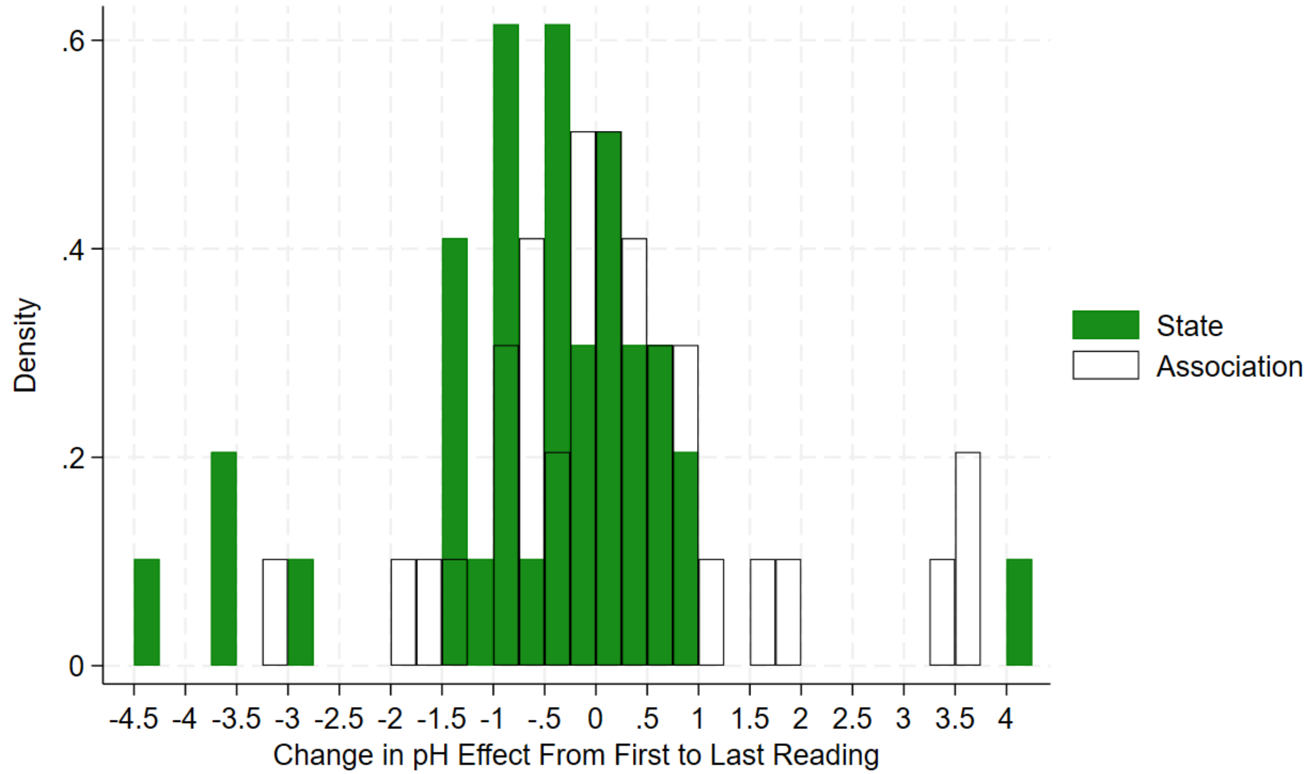

Figure A1: Histograms of State and Non-State Systems Effect on pH Over Time

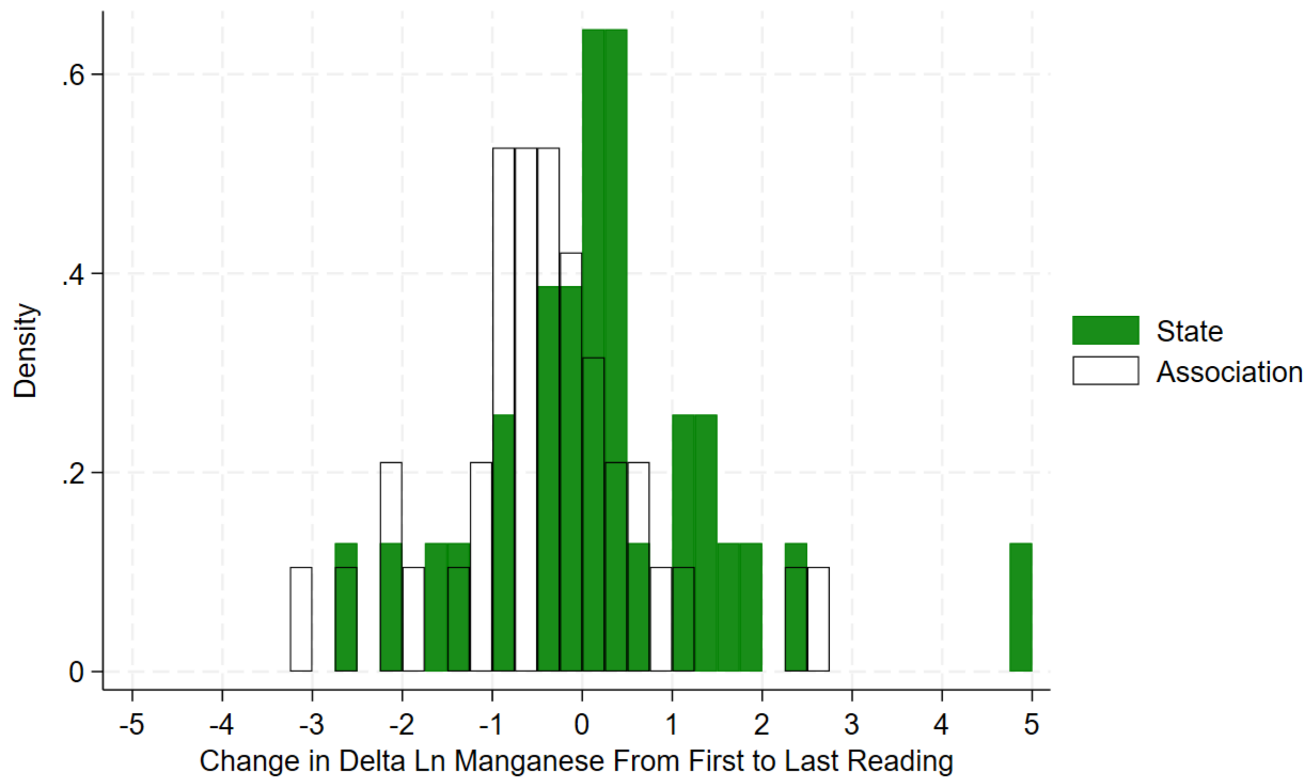

Figure A2: Histograms of State and Non-State Systems Effect on Manganese Over Time

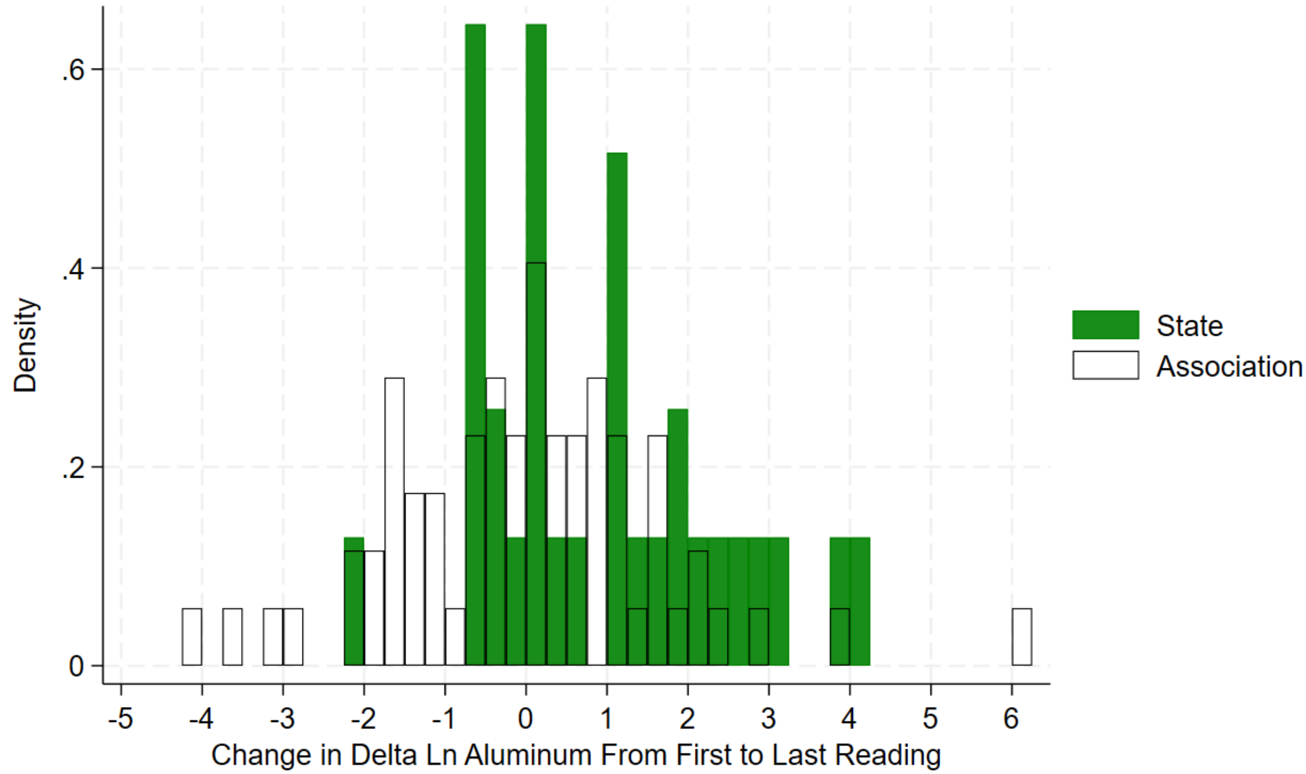

Figure A3: Histograms of State and Non-State Systems Effect on Aluminum Over Time

B Additional Tables: One-to-One Matching

Table B1: Descriptive Statistics: One-to-One Without Replacement MDM Sample

	Mean	SD	Min.	p25	p50	p75	Max	N
Characteristics	.	.	.	.	.	.	.	.
State System	0.50	0.50	0.00	0.00	0.50	1.00	1.00	102.00
Year Built	2,002.05	3.89	1,994.00	1,999.00	2,002.00	2,005.00	2,011.00	102.00
Number of Funding Sources	1.49	0.74	1.00	1.00	1.00	2.00	4.00	102.00
Severity	.	.	.	.	.	.	.	.
Initial Influent pH	4.03	1.21	2.50	3.10	3.52	4.45	7.21	78.00
Avg Influent pH	4.15	1.24	2.49	3.14	3.57	5.40	6.76	102.00
Influent Manganese (mg/l)	7.34	11.99	0.00	1.53	2.81	7.21	64.75	102.00
Influent Aluminum (mg/l)	12.67	21.58	0.00	0.56	5.04	12.75	144.62	102.00
Influent Iron (mg/l)	29.02	45.80	0.00	1.53	9.18	37.24	228.50	102.00
Influent TSS (mg/l)	12.32	16.96	0.00	5.00	6.45	14.00	118.61	102.00
Severity Threshold	.	.	.	.	.	.	.	.
Any Influent Standard Failed	0.99	0.10	0.00	1.00	1.00	1.00	1.00	102.00
Initial Influent pH ≤ 6	0.87	0.34	0.00	1.00	1.00	1.00	1.00	78.00
Avg Influent pH ≤ 6	0.92	0.27	0.00	1.00	1.00	1.00	1.00	102.00
Influent Mang. ≥ 2 (mg/l)	0.65	0.48	0.00	0.00	1.00	1.00	1.00	102.00
Influent Aluminum $\geq .5$ (mg/l)	0.77	0.42	0.00	1.00	1.00	1.00	1.00	102.00
Influent Iron ≥ 1.5 (mg/l)	0.75	0.43	0.00	1.00	1.00	1.00	1.00	102.00
Influent TSS ≥ 35 (mg/l)	0.05	0.22	0.00	0.00	0.00	0.00	1.00	102.00
Complexity	.	.	.	.	.	.	.	.
Number of Inflows	1.27	0.63	1.00	1.00	1.00	1.00	4.00	102.00
Size	.	.	.	.	.	.	.	.
Inflow Volume (gal/min)	290.18	737.97	2.04	18.89	52.67	184.33	5,637.25	102.00

Note: Data are from Datashed water quality reports and project documents.

Table B2: Descriptive Statistics: One-to-One Without Replacement MDM Sample, Outcome Variables

	Mean	SD	Min.	p25	p50	p75	Max	N
Effectiveness	.	.	.	.	.	.	.	.
Initial Change in pH	2.43	1.63	-2.10	0.80	2.80	3.90	4.95	78.00
Avg Change in pH	1.74	1.58	-1.97	0.50	1.35	3.21	4.72	102.00
Δ Ln Manganese (mg/l)	-0.67	1.21	-7.04	-0.96	-0.46	-0.12	2.79	101.00
Δ Ln Aluminum (mg/l)	-1.56	1.59	-6.86	-2.47	-1.33	-0.27	1.25	99.00
Δ Ln Iron (mg/l)	-1.84	1.88	-6.42	-3.04	-1.56	-0.50	3.86	101.00
Δ Ln TSS (mg/l)	0.08	0.87	-2.55	-0.46	0.11	0.60	2.36	97.00
Effectiveness Threshold	.	.	.	.	.	.	.	.
Any Effluent Standard Failed	0.74	0.44	0.00	0.00	1.00	1.00	1.00	102.00
Initial Effluent pH ≤ 6	0.21	0.41	0.00	0.00	0.00	0.00	1.00	78.00
Avg Effluent pH ≤ 6	0.42	0.50	0.00	0.00	0.00	1.00	1.00	102.00
Effluent Mang. ≥ 2 (mg/l)	0.43	0.50	0.00	0.00	0.00	1.00	1.00	102.00
Effluent Aluminum $\geq .5$ (mg/l)	0.49	0.50	0.00	0.00	0.00	1.00	1.00	102.00
Effluent Iron ≥ 1.5 (mg/l)	0.39	0.49	0.00	0.00	0.00	1.00	1.00	102.00
Effluent TSS ≥ 35 (mg/l)	0.05	0.22	0.00	0.00	0.00	0.00	1.00	102.00
Cost Effectiveness	.	.	.	.	.	.	.	.
Cost Per 100k Gallons Treated	144.23	198.40	0.51	20.85	78.92	186.43	1,152.23	101.00
Monitoring	.	.	.	.	.	.	.	.
Number of Readings	30.68	51.70	2.00	7.00	14.00	37.00	402.00	102.00
Maintenance	.	.	.	.	.	.	.	.
Change in pH Effectiveness	-0.18	1.43	-4.46	-0.82	-0.19	0.30	4.00	78.00
Change in Δ Ln Manganese	-0.10	1.31	-3.24	-0.78	-0.08	0.42	5.00	69.00
Change in Δ Ln Aluminum	0.45	1.42	-3.06	-0.56	0.21	1.36	4.20	67.00
Change in Δ Ln Iron	-0.02	1.84	-5.18	-1.04	0.22	1.11	4.37	68.00
Change in Δ Ln TSS	-0.12	1.32	-3.08	-0.92	-0.15	0.69	3.29	61.00

Note: Data are from Datashed water quality reports and project documents.

Table B3: Comparison of One-to-One Without Replacement MDM Matched State and Association Systems

	Association System (N=51)	State System (N=51)	Diff.	Normalized Diff.	P-Value
Characteristics	.	.	.	.	.
Year Built	2,002.84	2,001.25	1.59	0.42	0.04
Number of Funding Sources	1.45	1.53	-0.08	-0.11	0.60
Severity	.	.	.	.	.
Initial Influent pH	4.07	4.00	0.06	0.05	0.82
Avg Influent pH	4.16	4.13	0.03	0.03	0.90
Influent Manganese (mg/l)	6.58	8.10	-1.52	-0.13	0.53
Influent Aluminum (mg/l)	10.42	14.92	-4.50	-0.21	0.30
Influent Iron (mg/l)	24.38	33.65	-9.28	-0.20	0.31
Influent TSS (mg/l)	11.56	13.08	-1.51	-0.09	0.66
Severity Threshold	.	.	.	.	.
Any Influent Standard Failed	0.98	1.00	-0.02	-0.20	0.32
Initial Influent pH ≤ 6	0.87	0.87	0.00	0.00	1.00
Avg Influent pH ≤ 6	0.92	0.92	0.00	0.00	1.00
Influent Mang. ≥ 2 (mg/l)	0.63	0.67	-0.04	-0.08	0.68
Influent Aluminum $\geq .5$ (mg/l)	0.82	0.73	0.10	0.23	0.24
Influent Iron ≥ 1.5 (mg/l)	0.73	0.78	-0.06	-0.14	0.49
Influent TSS ≥ 35 (mg/l)	0.04	0.06	-0.02	-0.09	0.65
Complexity	.	.	.	.	.
Number of Inflows	1.25	1.29	-0.04	-0.06	0.76
Size	.	.	.	.	.
Inflow Volume (gal/min)	302.03	278.34	23.69	0.03	0.87

Note: Normalized Mean Differences are calculated using the method of Imbens and Wooldridge (2009), which is the difference in means divided by the square root of the average variance. P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Severity Threshold variables are all binary variables.

Table B4: Mean Differences: One-to-One Without Replacement MDM Matched State and Association Systems

	Association System (N=51)	State System (N=51)	Diff.	P-Value
Effectiveness	.	.	.	.
Initial Change in pH	2.14	2.71	-0.57	0.12
Avg Change in pH	1.76	1.71	0.05	0.87
Δ Ln Manganese (mg/l)	-0.68	-0.66	-0.02	0.95
Δ Ln Aluminum (mg/l)	-1.66	-1.46	-0.21	0.52
Δ Ln Iron (mg/l)	-1.56	-2.12	0.56	0.14
Δ Ln TSS (mg/l)	0.05	0.11	-0.06	0.75
Effectiveness Threshold	.	.	.	.
Any Effluent Standard Failed	0.67	0.80	-0.14	0.12
Initial Effluent pH ≤ 6	0.26	0.15	0.10	0.27
Avg Effluent pH ≤ 6	0.37	0.47	-0.10	0.32
Effluent Mang. ≥ 2 (mg/l)	0.37	0.49	-0.12	0.23
Effluent Aluminum $\geq .5$ (mg/l)	0.39	0.59	-0.20	0.05
Effluent Iron ≥ 1.5 (mg/l)	0.39	0.39	0.00	1.00
Effluent TSS ≥ 35 (mg/l)	0.06	0.04	0.02	0.65
Cost Effectiveness	.	.	.	.
Ln Cost 100k Gallons Treated	3.83	4.34	-0.52	0.09
Monitoring	.	.	.	.
Ln Number of Readings	2.48	2.93	-0.45	0.05
Maintenance	.	.	.	.
Change in pH Effectiveness	0.15	-0.52	0.67	0.04
Change in Δ Ln Manganese	-0.41	0.28	-0.69	0.03
Change in Δ Ln Aluminum	0.06	0.89	-0.83	0.02
Change in Δ Ln Iron	-0.20	0.20	-0.41	0.38
Change in Δ Ln TSS	-0.20	-0.02	-0.18	0.60

Note: P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Effectiveness Threshold variables are all binary variables.

Table B5: Biased Corrected Effects: Effectiveness, One-to-One Without Replacement

	Any	Initial pH		pH		Mn		Al		Fe	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	p(Fail)	Δ	p(Fail)	Δ	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)
State System	0.29**	0.16	-0.06	-0.70	0.21	0.37	0.07	0.57	0.40**	0.27	0.19
	(0.11)	(0.39)	(0.13)	(0.37)	(0.15)	(0.26)	(0.18)	(0.39)	(0.13)	(0.42)	(0.17)
Ln Influent Flow (gal/min)	0.05	0.08	-0.03	-0.07	0.02	0.16	0.10	0.16	0.02	0.18	0.06
	(0.03)	(0.11)	(0.03)	(0.10)	(0.04)	(0.09)	(0.06)	(0.09)	(0.03)	(0.10)	(0.05)
Ln Number of Inflows	-0.24	0.07	-0.05	0.42	-0.20	-0.95**	-0.06	-0.89	-0.36*	-1.40*	-0.05
	(0.14)	(0.53)	(0.16)	(0.50)	(0.16)	(0.32)	(0.18)	(0.57)	(0.14)	(0.54)	(0.14)
R-squared	0.26	0.28	0.20	0.21	0.12	0.30	0.19	0.22	0.30	0.26	0.26
N	98	75	64	99	91	98	61	96	77	98	73
Cond. on Failed Influent Std.	Y	N	Y	N	Y	N	Y	N	Y	N	Y

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. Each system in the regressions is weighted using analytical weights of the number of readings that contribute its average pH and pollutant concentrations.

Table B6: Biased Corrected Effects: Cost-Effectiveness and Monitoring, One-to-One Without Replacement

	Ln(Cost Per 100k Gallons)	Ln(Number of Readings)
	(1)	(2)
State System	0.58	0.48
	(0.34)	(0.28)
Ln Number of Inflows	1.02**	0.49
	(0.38)	(0.32)
Ln Influent Flow (gal/min)		0.08
		(0.08)
R-squared	0.20	0.27
N	97	99

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction.

Table B7: Biased Corrected Effects: Maintenance, One-to-One Without Replacement

	Change in Δ pH	Change in Δ Ln			
	(1)	(2)	(3)	(4)	(5)
	pH	Mn	Al	Fe	TSS
State System	-0.70	0.69	0.61	0.07	0.41
	(0.39)	(0.46)	(0.39)	(0.66)	(0.55)
Ln Influent Flow (gal/min)	0.07	-0.08	0.12	0.10	-0.02
	(0.08)	(0.11)	(0.14)	(0.21)	(0.18)
Ln Number of Inflows	-0.55	0.55	0.80*	1.35*	0.35
	(0.38)	(0.33)	(0.36)	(0.53)	(0.43)
Ln Measurement Duration	-0.58*	0.29	0.32	0.38	-0.02
	(0.26)	(0.46)	(0.26)	(0.47)	(0.36)
R-squared	0.29	0.18	0.33	0.23	0.16
N	75	65	62	64	56

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction.

C Additional Tables: One-to-Two Matching With Replacement

Table C1: Descriptive Statistics: One-to-Two With Replacement MDM Sample

	Mean	SD	Min.	p25	p50	p75	Max	N
Characteristics	.	.	.	.	.	.	.	.
State System	0.33	0.47	0.00	0.00	0.00	1.00	1.00	153.00
Year Built	2,001.92	3.96	1,992.00	1,999.00	2,002.00	2,005.00	2,012.00	153.00
Number of Funding Sources	1.46	0.72	1.00	1.00	1.00	2.00	4.00	153.00
Severity	.	.	.	.	.	.	.	.
Initial Influent pH	4.10	1.23	2.50	3.10	3.65	4.70	7.21	117.00
Avg Influent pH	4.15	1.18	2.49	3.19	3.60	5.34	6.76	153.00
Influent Manganese (mg/l)	6.42	11.13	0.00	1.52	2.49	6.42	64.75	153.00
Influent Aluminum (mg/l)	11.11	18.68	0.00	0.93	6.02	11.62	144.62	153.00
Influent Iron (mg/l)	24.61	43.87	0.00	0.92	5.96	28.18	228.50	153.00
Influent TSS (mg/l)	12.39	15.52	0.00	5.00	6.60	14.00	118.61	153.00
Severity Threshold	.	.	.	.	.	.	.	.
Any Influent Standard Failed	0.99	0.11	0.00	1.00	1.00	1.00	1.00	153.00
Initial Influent pH ≤ 6	0.84	0.37	0.00	1.00	1.00	1.00	1.00	117.00
Avg Influent pH ≤ 6	0.92	0.27	0.00	1.00	1.00	1.00	1.00	153.00
Influent Mang. ≥ 2 (mg/l)	0.63	0.49	0.00	0.00	1.00	1.00	1.00	153.00
Influent Aluminum $\geq .5$ (mg/l)	0.80	0.40	0.00	1.00	1.00	1.00	1.00	153.00
Influent Iron ≥ 1.5 (mg/l)	0.71	0.45	0.00	0.00	1.00	1.00	1.00	153.00
Influent TSS ≥ 35 (mg/l)	0.06	0.24	0.00	0.00	0.00	0.00	1.00	153.00
Complexity	.	.	.	.	.	.	.	.
Number of Inflows	1.24	0.58	1.00	1.00	1.00	1.00	4.00	153.00
Size	.	.	.	.	.	.	.	.
Inflow Volume (gal/min)	249.15	638.70	1.75	19.39	52.83	162.50	5,637.25	153.00

Note: Data are from Datashed water quality reports and project documents.

Table C2: Descriptive Statistics: One-to-Two With Replacement MDM Sample, Outcome Variables

	Mean	SD	Min.	p25	p50	p75	Max	N
Effectiveness	.	.	.	.	.	.	.	.
Initial Change in pH	2.32	1.69	-2.10	0.78	2.80	3.90	4.95	117.00
Avg Change in pH	1.79	1.54	-1.97	0.56	1.49	3.20	4.72	153.00
Δ Ln Manganese (mg/l)	-0.70	1.09	-7.04	-0.96	-0.46	-0.17	2.79	151.00
Δ Ln Aluminum (mg/l)	-1.64	1.55	-6.86	-2.56	-1.36	-0.53	1.25	148.00
Δ Ln Iron (mg/l)	-1.69	1.85	-6.42	-3.14	-1.35	-0.50	3.86	151.00
Δ Ln TSS (mg/l)	0.03	0.86	-2.55	-0.46	0.04	0.58	2.36	147.00
Effectiveness Threshold	.	.	.	.	.	.	.	.
Any Effluent Standard Failed	0.71	0.46	0.00	0.00	1.00	1.00	1.00	153.00
Initial Effluent pH ≤ 6	0.21	0.41	0.00	0.00	0.00	0.00	1.00	117.00
Avg Effluent pH ≤ 6	0.41	0.49	0.00	0.00	0.00	1.00	1.00	153.00
Effluent Mang. ≥ 2 (mg/l)	0.37	0.48	0.00	0.00	0.00	1.00	1.00	153.00
Effluent Aluminum $\geq .5$ (mg/l)	0.46	0.50	0.00	0.00	0.00	1.00	1.00	153.00
Effluent Iron ≥ 1.5 (mg/l)	0.37	0.48	0.00	0.00	0.00	1.00	1.00	153.00
Effluent TSS ≥ 35 (mg/l)	0.05	0.21	0.00	0.00	0.00	0.00	1.00	153.00
Cost Effectiveness	.	.	.	.	.	.	.	.
Cost Per 100k Gallons Treated	133.15	180.53	0.15	20.85	76.96	187.86	1,152.23	152.00
Monitoring	.	.	.	.	.	.	.	.
Number of Readings	28.83	45.50	2.00	6.00	14.00	36.00	402.00	153.00
Maintenance	.	.	.	.	.	.	.	.
Change in pH Effectiveness	-0.01	1.44	-4.46	-0.60	0.00	0.52	4.00	117.00
Change in Δ Ln Manganese	-0.19	1.22	-3.24	-0.77	-0.31	0.33	5.00	106.00
Change in Δ Ln Aluminum	0.42	1.47	-3.06	-0.56	0.21	1.36	6.11	102.00
Change in Δ Ln Iron	0.03	1.70	-5.18	-1.04	0.22	1.11	4.37	104.00
Change in Δ Ln TSS	-0.17	1.33	-3.08	-1.16	-0.15	0.68	3.29	94.00

Note: Data are from Datashed water quality reports and project documents.

Table C3: Comparison of One-to-Two With Replacement MDM Matched State and Association Constructed Systems

	Association System (N=102)	State System (N=51)	Diff.	Normalized Diff.	P-Value
Characteristics	.	.	.	.	.
Year Built	2,002.25	2,001.25	0.99	0.25	0.14
Number of Funding Sources	1.43	1.53	-0.10	-0.13	0.45
Severity	.	.	.	.	.
Initial Influent pH	4.15	4.00	0.15	0.12	0.55
Avg Influent pH	4.16	4.13	0.03	0.03	0.88
Influent Manganese (mg/l)	5.58	8.10	-2.51	-0.22	0.21
Influent Aluminum (mg/l)	9.21	14.92	-5.70	-0.27	0.16
Influent Iron (mg/l)	20.08	33.65	-13.57	-0.30	0.09
Influent TSS (mg/l)	12.04	13.08	-1.03	-0.06	0.74
Severity Threshold	.	.	.	.	.
Any Influent Standard Failed	0.98	1.00	-0.02	-0.20	0.16
Initial Influent pH ≤ 6	0.82	0.87	-0.05	-0.14	0.46
Avg Influent pH ≤ 6	0.92	0.92	0.00	0.00	1.00
Influent Mang. ≥ 2 (mg/l)	0.61	0.67	-0.06	-0.12	0.48
Influent Aluminum $\geq .5$ (mg/l)	0.83	0.73	0.11	0.26	0.14
Influent Iron ≥ 1.5 (mg/l)	0.68	0.78	-0.11	-0.24	0.15
Influent TSS ≥ 35 (mg/l)	0.06	0.06	0.00	0.00	1.00
Complexity	.	.	.	.	.
Number of Inflows	1.21	1.29	-0.09	-0.15	0.40
Size	.	.	.	.	.
Inflow Volume (gal/min)	234.55	278.34	-43.79	-0.07	0.68

Note: Normalized mean differences are calculated using the method of Imbens and Wooldridge (2009), which is the difference in means divided by the square root of the average variance. P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Severity Threshold variables are all binary variables.

Table C4: Mean Differences: One-to-Two With Replacement MDM Matched State and Association Constructed Systems

	Association System (N=102)	State System (N=51)	Diff.	P-Value
Effectiveness	.	.	.	.
Initial Change in pH	2.12	2.71	-0.60	0.06
Avg Change in pH	1.83	1.71	0.11	0.67
Δ Ln Manganese (mg/l)	-0.73	-0.66	-0.07	0.73
Δ Ln Aluminum (mg/l)	-1.73	-1.46	-0.28	0.31
Δ Ln Iron (mg/l)	-1.48	-2.12	0.64	0.04
Δ Ln TSS (mg/l)	0.00	0.11	-0.11	0.53
Effectiveness Threshold	.	.	.	.
Any Effluent Standard Failed	0.66	0.80	-0.15	0.05
Initial Effluent pH $\leq$ 6	0.23	0.15	0.08	0.31
Avg Effluent pH $\leq$ 6	0.38	0.47	-0.09	0.30
Effluent Mang. $\geq$ 2 (mg/l)	0.30	0.49	-0.19	0.03
Effluent Aluminum $\geq$ .5 (mg/l)	0.40	0.59	-0.19	0.03
Effluent Iron $\geq$ 1.5 (mg/l)	0.35	0.39	-0.04	0.64
Effluent TSS $\geq$ 35 (mg/l)	0.05	0.04	0.01	0.78
Cost Effectiveness	.	.	.	.
Ln Cost 100 Gallons Treated	3.87	4.34	-0.48	0.08
Monitoring	.	.	.	.
Ln Number of Readings	2.55	2.93	-0.39	0.06
Maintenance	.	.	.	.
Change in pH Effectiveness	0.25	-0.52	0.76	0.01
Change in Δ Ln Manganese	-0.38	0.28	-0.66	0.02
Change in Δ Ln Aluminum	0.21	0.89	-0.68	0.04
Change in Δ Ln Iron	-0.04	0.20	-0.25	0.56
Change in Δ Ln TSS	-0.22	-0.02	-0.21	0.54

Note: P-Values are two-tailed and from a t-test for the difference in means of unpaired data, assuming unequal variances. The Effectiveness Threshold variables are all binary variables.

Table C5: Biased Corrected Effects: Effectiveness, One-to-Two Matching With Replacement

	Any	Initial pH		pH		Mn		Al		Fe	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	p(Fail)	Δ	p(Fail)	Δ	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)	Δ Ln	p(Fail)
State System	0.16	0.53	-0.12	-0.38	0.16	0.25	0.10	0.40	0.32**	-0.18	0.15
	(0.08)	(0.37)	(0.12)	(0.29)	(0.11)	(0.19)	(0.13)	(0.31)	(0.10)	(0.36)	(0.11)
Ln Influent Flow (gal/min)	0.06**	0.10	-0.02	-0.09	0.05	0.12*	0.08	0.14	0.02	0.05	0.08
	(0.02)	(0.11)	(0.03)	(0.08)	(0.03)	(0.05)	(0.05)	(0.07)	(0.03)	(0.09)	(0.04)
Ln Number of Inflows	0.01	-0.01	-0.01	0.43	-0.21	-0.55*	0.13	-0.92	-0.27	-1.29**	0.12
	(0.15)	(0.49)	(0.16)	(0.47)	(0.14)	(0.27)	(0.16)	(0.50)	(0.15)	(0.47)	(0.13)
R-squared	0.19	0.31	0.21	0.21	0.11	0.23	0.24	0.26	0.27	0.25	0.27
N	147	113	93	149	135	148	89	145	117	148	104
Cond. on Failed Influent Std.	Y	N	Y	N	Y	N	Y	N	Y	N	Y

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. Using frequency weights, we weight each state system in the regressions by one, and the 65 association systems by the number of times they are matched to a state system.

Table C6: Biased Corrected Effects: Cost-Effectiveness and Monitoring, One-to-Two Matching With Replacement

	Ln(Cost Per 100k Gallons)	Ln(Number of Readings)
	(1)	(2)
State System	0.60*	0.41
	(0.30)	(0.22)
Ln Number of Inflows	0.76*	0.73*
	(0.37)	(0.29)
Ln Influent Flow (gal/min)		0.09
		(0.07)
R-squared	0.16	0.26
N	147	149

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. Using frequency weights, we weight each state system in the regressions by one, and the 65 association systems by the number of times they are matched to a state system.

Table C7: Biased Corrected Effects: Maintenance, One-to-Two Matching With Replacement

	Change in Δ pH	Change in Δ Ln			
	(1)	(2)	(3)	(4)	(5)
	pH	Mn	Al	Fe	TSS
State System	-0.93**	0.66	0.66*	-0.08	0.26
	(0.32)	(0.37)	(0.30)	(0.51)	(0.46)
Ln Influent Flow (gal/min)	0.06	-0.05	0.04	0.12	0.01
	(0.07)	(0.08)	(0.10)	(0.15)	(0.15)
Ln Number of Inflows	-0.55	0.43	0.41	0.86	0.16
	(0.32)	(0.34)	(0.27)	(0.45)	(0.41)
Ln Measurement Duration	-0.37*	-0.06	0.15	0.09	-0.07
	(0.18)	(0.32)	(0.18)	(0.34)	(0.28)
R-squared	0.35	0.19	0.37	0.22	0.13
N	113	102	96	100	88

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. All models include fixed effects for the year of system construction. Using frequency weights, we weight each state system in the regressions by one, and the 65 association systems by the number of times they are matched to a state system.