

Operational Service Conditions are the primary predictors of Household Water Service Quality Across Heterogeneous Provision Systems

Aadhitya Mohanavelu

Stanford University

Khalid Osman

osmank@stanford.edu

Stanford University

Article

Keywords: Water service quality, water infrastructure, household water security, heterogeneous water provision, service delivery, water systems, user experience, priority-weighted measurement, machine learning, SHAP

Posted Date: September 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10888037/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Operational Service Conditions are the primary predictors of Household Water Service Quality Across Heterogeneous Provision Systems

Aadhitya Mohanavelu¹, Khalid Osman^{1,2}

¹ Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA

² Corresponding Author: osmank@stanford.edu

Abstract

Conventional source- or nominal access-based indicators may overlook heterogeneity in the water-service dimensions that shape everyday household experience, including reliability, affordability, accessibility, quantity, acceptability, perceived quality, responsiveness, and household effort. We developed a household-level Water Service Quality Score (WSQS) using 6,145 household surveys from South India. WSQS integrates eight service domains and compares equal-, population-priority-, and individual-priority-weighted formulations. We modeled WSQS using service-support, household-context, and stated-priority predictors, evaluated performance across multiple predictive models, and interpreted predictions using SHAP attribution and nonlinear service-response analyses. Within the fitted model, service-support conditions accounted for most of the predictable variation and 94.9% of grouped SHAP attribution. Important predictors included value for money, complaint resolution, water pressure, appearance acceptability, water clarity, provider contact, and shortage experience. Household priorities contributed modest, domain-specific heterogeneity but little incremental predictive performance. These findings show that household water service quality is multidimensional and measurable beyond access, helping water-sector decision-makers identify and prioritize service deficiencies across heterogeneous provision systems.

Keywords: Water service quality; water infrastructure; household water security; heterogeneous water provision; service delivery, water systems; user experience; priority-weighted measurement; machine learning; SHAP

Introduction

Domestic water service in many low- and middle-income countries is rarely experienced through a single, stable infrastructure pathway (Daly et al., 2021; Elliott et al. 2019). Although households are often classified by their main or improved water source, everyday water use may depend on overlapping arrangements. These sources may include piped supply, groundwater self-supply, tanker delivery, commercial water, shared sources, rainwater harvesting, and other supplementary systems. A household may use one source for drinking, another for cooking or washing, and shift between sources when supply becomes intermittent, costly, insufficient, or mistrusted. This complicates monitoring because a primary-source category can conceal the multiple arrangements households use to meet different needs. Systematic evidence from low- and middle- income countries (LMIC) shows that multiple-source use is common, including among households with improved primary sources, and unreliable supply can trigger storage, treatment, source switching, and other coping strategies that reshape household service experience (Daly et al., 2021; Lee et al., 2020). In these settings, source type is an incomplete proxy for service quality; for example, service level may be measured through an established piped supply connection (Grafton et al., 2025; Geleijnse et al., 2023). However, the same nominal source can produce very different experiences of reliability, affordability, sufficiency, acceptability, and accessibility.

This creates a measurement problem for water-service monitoring. Conventional indicators, including improved-source classifications, piped-connection status, and service ladders, remain essential for comparison and planning, but they compress distinct dimensions of service into infrastructure-based categories (Kayser et al., 2013; Vilarinho et al., 2024). Research on domestic water-service delivery has long emphasized the need to monitor safety, quantity, accessibility, reliability, and cost rather than source type alone (Kayser et al., 2013; Vilarinho et al., 2024; Hannah et al., 2020). Recent global mapping further shows that even the use of safely managed drinking water remains data-limited across many LMICs

(Greenwood et al., 2024). At the household level, access and service quality can also diverge sharply; evidence from Bangalore, India, shows that public-service access and piped-water service quality are distinct outcomes shaped by institutional and political processes (Kumar et al., 2022). Monitoring infrastructure access is therefore necessary, but it is not equivalent to measuring end-user service quality.

Household surveys provide one way to measure service dimensions that are often invisible in infrastructure records, utility reports, or source classifications. They can capture pressure, interruptions, shortages, cost burden, waiting time, perceived quality, acceptability, collection difficulty, satisfaction, and provider responsiveness (Bazaanah et al., 2024; Osman and Faust, 2021; Hopkins et al., 2024; Hamilton et al., 2024). Experience-based water insecurity research has shown that household water problems are multidimensional and measurable across diverse LMIC contexts (Young et al., 2019). Related methodological work emphasizes that micro-level water metrics must address dimensionality, temporality, unit of analysis, and comparability, while also accounting for capabilities, institutions, governance, and socio-cultural dynamics (Jepson et al., 2017; Wutich et al., 2017). Yet converting household survey responses into a service-quality metric remains difficult. Survey items typically span physical, experiential, institutional, and evaluative dimensions; they may operate on different ordinal scales; and they do not necessarily contribute equally to overall service experience. A single satisfaction question may summarize a household's judgment, but it cannot diagnose whether poor service reflects unreliable supply, low pressure, unaffordable cost, insufficient quantity, unacceptable quality, access burden, or weak customer support.

Aggregation also raises a preference problem. A metric that combines affordability, access, reliability, efficiency, acceptability, perceived quality, quantity sufficiency, and customer service must decide how to weight those domains. Equal weighting provides a transparent baseline, but households may value service dimensions differently. Stated-preference studies show heterogeneity in how households value improvements in water quality, reliability, pressure, system management, and drinkability, while customer satisfaction studies show that perceived health risk, organoleptic characteristics, payment systems, and price shape service judgments (Rajapakshe et al., 2023; Vásquez et al., 2021; Denantes and Donoso, 2021; Ogata et al., 2026). These findings suggest that priorities may shape how service conditions are interpreted, weighted, and translated into overall service-quality judgments. At the same time, embedding stated priorities directly into the modeled outcome can make later priority-based prediction circular. The measurement challenge is therefore to retain a transparent and comparable service-quality metric while still allowing household priorities to be examined as a source of heterogeneity in service experience.

Existing work has advanced several parts of this problem. Global monitoring studies provide standardized indicators; household water insecurity scales measure lived water problems; multiple-source and coping studies show how households respond to unreliable provision; satisfaction studies identify perceived determinants of service experience; and customer-centered utility indices formalize service-quality assessment at the provider level (Young et al., 2019; Daly et al., 2021; Lee et al., 2020; Denantes and Donoso, 2021; Vilarinho et al., 2024; Skerker et al., 2026). Recent monitoring studies have also argued for moving beyond isolated measurements, challenged narrow quantity-versus-quality framings, and linked household water outcomes to tap-water service conditions (Charles et al., 2020; Nowicki et al., 2020; Jeandron et al., 2019). What remains limited is an integrated household-level framework that combines multiple domains of end-user service quality, compares priority-sensitive aggregation, and tests whether household priorities add explanatory value beyond observed service conditions and household context.

We address this gap using survey data from 6,145 households in South India, primarily Tamil Nadu, where households rely on heterogeneous water arrangements including piped supply, groundwater, tanker delivery, and commercial water (Fig. 1). The Water Service Quality Score (WSQS) combines eight CORE-8 domains: affordability, access, reliability, efficiency, acceptability, perceived quality, quantity sufficiency, and customer service. We compare equal-weight, population-priority, and individual-priority weighting architectures; assess item-level similarity using correlation and exploratory factor analysis; model equal-weight WSQS using nested predictor sets; and evaluate whether household service quality is

explained primarily by operational service-support conditions, household context, or stated priorities. Flexible predictive models are used diagnostically (not causally), and interpreted using SHAP attribution, nonlinear response analysis, priority-conditioned contrasts, and robustness checks across outcome definitions, model classes, context specifications, and repeated train-test splits.

The contribution of this paper is a household-level framework for measuring and diagnosing end-user water service quality in heterogeneous provision systems. Existing household water-insecurity measures, such as HWISE, quantify the severity of water-insecurity experiences (Young et al., 2019). WSQS instead evaluates service performance and identifies the service dimensions in which deficiencies occur. The framework moves beyond source type, infrastructure access, and single satisfaction measures; distinguishes baseline measurement from priority-sensitive interpretation; tests whether priorities add information beyond observed service conditions and household context; and identifies the physical, experiential, and institutional conditions most strongly associated with household water service quality. It connects service-quality measurement to broader efforts in water-service monitoring, water security, and equity-oriented infrastructure assessment (Charles et al., 2020; Osman and Faust, 2021; Mohanavelu et al., 2024; Mohanavelu and Osman, 2024; Savelli et al., 2023), while reframing monitoring from classifying access to diagnosing how service functions in household experience.

Study area and household water provision types

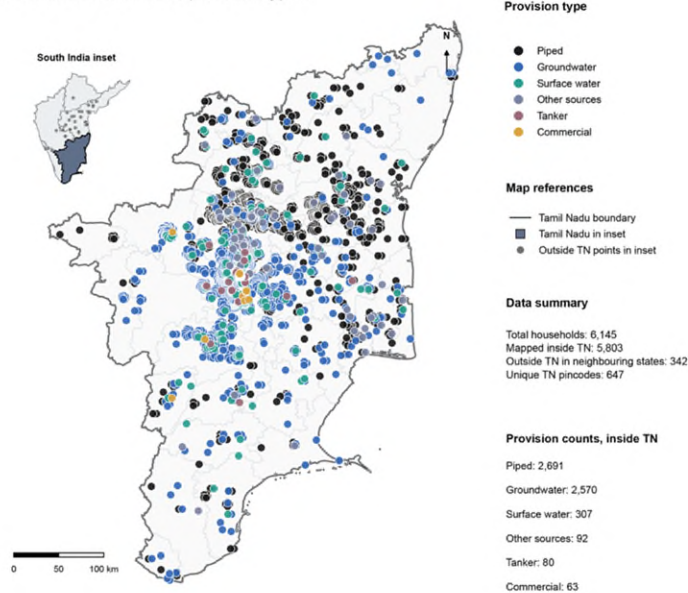


Figure 1. Study area and household water provision types used for analysis.

3. Results

Multidimensional WSQS captures heterogeneity across service domains

The final WSQS construction dataset included 6,145 households. The Water Service Quality Score (WSQS) was constructed from eight service-quality domains: affordability, access, reliability, efficiency, acceptability, perceived quality, quantity sufficiency, and customer service. Each domain was harmonized on a common direction-checked service-quality scale and transformed to the 0–100 reporting scale, where higher values indicate better service quality (Fig. 2a). The domain response distributions showed substantial

variation across service dimensions, supporting the need for a multidimensional measure rather than a single source-type, access, or satisfaction indicator (Fig. 2b; Table 1).

Mean domain scores ranged from 64.93 for acceptability to 76.08 for reliability on the 0–100 reporting scale (Table 1). Quantity, access, and perceived quality also scored relatively high, with means of 75.52, 75.24, and 72.96, respectively. Efficiency was intermediate, with a mean of 71.22, while affordability and customer service were both lower, with means of 68.83. This pattern indicates that households may report relatively better reliability, quantity, access, or perceived quality, while still facing burdens in acceptability, affordability, or customer service. These domain-level differences introduce the service-function perspective developed throughout the paper. Rather than treating access or source type as a sufficient summary, WSQS asks how well water service functions across the dimensions households experience.

Respondent-stated priorities also varied across domains (Fig. 2c; Table 2). These priorities were retained on the original 0–10 scale and used to evaluate whether household-stated domain importance materially changed WSQS measurement. The purpose of this comparison was not to replace the primary score with a preference-weighted outcome, but to test whether priority weighting meaningfully changed the aggregate distribution or individual household scores.

The three WSQS weighting architectures produced similar aggregate distributions. The equal-weight, population-priority, and individual-priority WSQS means were 71.74, 71.74, and 71.77, respectively, on the 0–100 reporting scale (Table 1). Individual-priority weighting shifted scores for some households, but the aggregate shift was limited: the mean shift was 0.02 points, the median shift was 0.00, the mean absolute shift was 0.95, the 95th percentile absolute shift was 3.71, and the correlation between equal-weight and individual-priority WSQS was 0.991 (Fig. 2d; Table 3). These results support retaining equal-weight WSQS as the primary non-circular modeling outcome, with population-priority and individual-priority WSQS treated as sensitivity outcomes.

Multidimensional WSQS captures heterogeneity across service domains

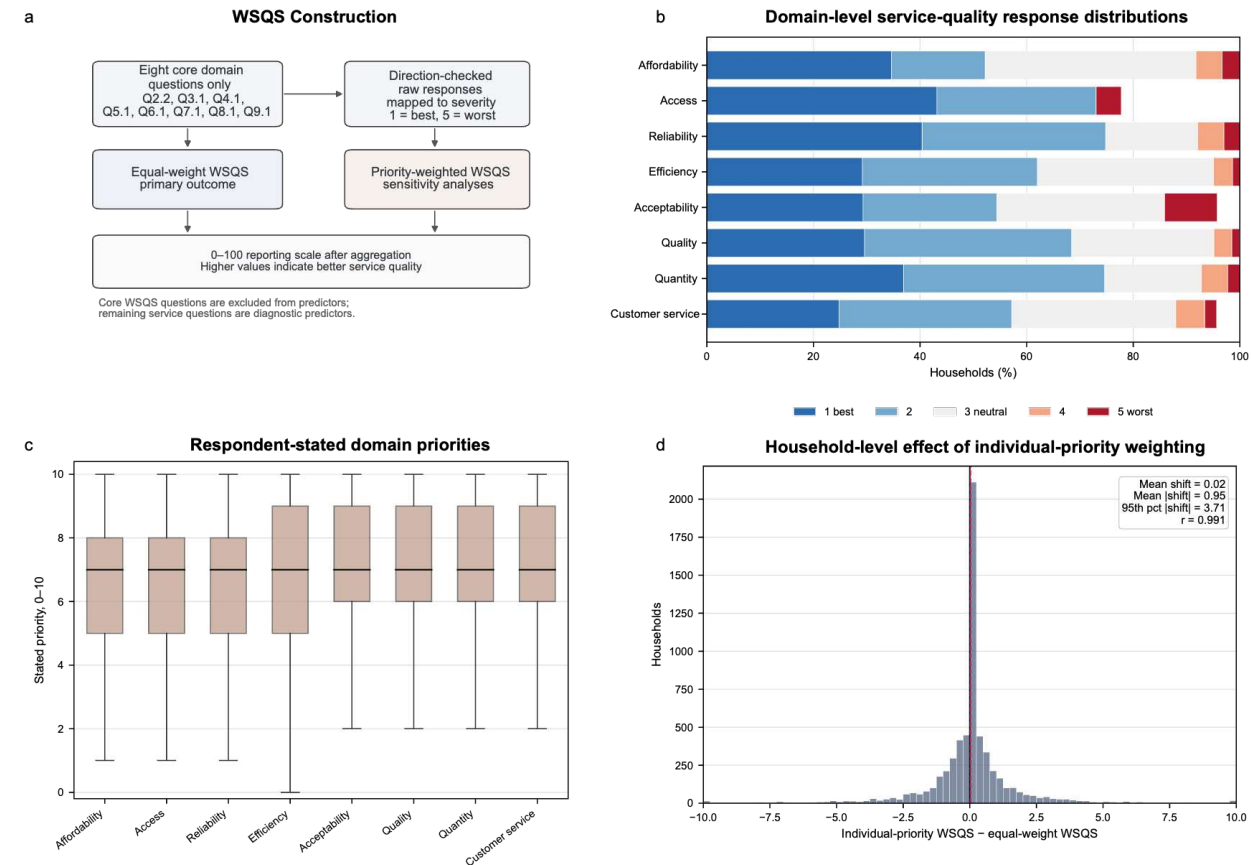


Figure 2. Multidimensional CORE-8 WSQS captures heterogeneity across service domains and stated priorities. a) Domain response structuring to constructs for predictive modeling. b) Domain-level service-quality response distributions across the eight CORE-8 WSQS domains after direction-checked mapping to a common 1–5 severity scale. c) Respondent-stated priorities across the eight WSQS domains on the original 0–10 scale, with all household observations shown. d) Household-level difference between individual-priority-weighted and equal-weight WSQS on the 0–100 reporting scale. Higher WSQS values indicate better service quality.

Table 1. CORE-8 domain and aggregate WSQS scores.

Measure	Mean	SD
Affordability domain	68.83	27.10
Access domain	75.24	27.99
Reliability domain	76.08	25.50
Efficiency domain	71.22	23.25
Acceptability domain	64.93	30.67
Quality domain	72.96	22.65

Quantity domain	75.52	24.39
Customer service domain	68.83	24.48
Equal-weight WSQS	71.74	15.79
Population-priority WSQS	71.74	15.81
Individual-priority WSQS	71.77	15.93

Table 2. Priority-weight summary by WSQS domain.

Domain	Mean stated priority	SD	Median	P25	P75	Population-priority weight
Affordability	6.70	2.26	7.00	5.00	8.00	0.120
Access	6.94	2.14	7.00	5.00	8.00	0.124
Reliability	6.90	2.14	7.00	5.00	8.00	0.124
Efficiency	6.92	2.16	7.00	5.00	9.00	0.124
Acceptability	7.06	2.13	7.00	6.00	9.00	0.126
Quality	7.07	2.11	7.00	6.00	9.00	0.127
Quantity	7.12	2.08	7.00	6.00	9.00	0.128
Customer service	7.12	2.13	7.00	6.00	9.00	0.127

WSQS structure is coherent across domains and stable across weighting schemes

The domain correlation, quintile-profile, and weighting-robustness analyses were used to evaluate complementary aspects of WSQS structure. Domain correlations tested whether the eight WSQS domains were related enough to support aggregation, but not so highly correlated that they were redundant. Quintile profiles tested whether the aggregate equal-weight WSQS ordered households coherently across the underlying domains. Weighting comparisons tested whether the aggregate score was sensitive to equal, population-priority, or individual-priority weighting.

Correlations among domain scores were generally positive, indicating that households reporting better service in one domain also tended to report better service in other domains. However, the correlations were moderate rather than near perfect. The correlation matrix showed values ranging from 0.12 to 0.52 outside the diagonal, with the highest pairwise correlation observed between perceived quality and customer service (Fig. 3a). This supports the interpretation of WSQS as a multidimensional index; the domains are related, but each contributes distinct information. Supplementary measurement-structure diagnostics were consistent with this interpretation. Exploratory factor analysis of the broader 32-item service-quality set identified six empirical response factors rather than a one-to-one reproduction of the eight conceptually specified WSQS domains. This is expected because the WSQS domains were specified as service-quality dimensions for measurement and diagnosis, whereas the EFA was used descriptively to examine how

broader survey responses clustered empirically. The eight CORE-8 construction indicators nevertheless showed moderate cross-domain coherence (Cronbach's $\alpha = 0.757$; Supplementary Information, Section S1.3). Together, these results support treating WSQS as a multidimensional composite of related but non-interchangeable service dimensions rather than as a unidimensional latent scale.

The quintile-profile analysis provided a second structural check. If the aggregate WSQS is meaningful, households in higher WSQS quintiles should generally show better scores across the component domains. This pattern was observed. The lowest equal-weight WSQS quintile had consistently lower domain scores, while the highest quintile had substantially higher scores across nearly all domains (Fig. 3b). For example, acceptability increased from 45.8 in the lowest quintile to 95.0 in the highest quintile, and customer service increased from 45.9 to 93.4. The stronger gradients in the quintile-profile heatmap, compared with the more moderate domain-correlation matrix, are informative as well. They indicate that domains are not redundant at the item level, but they still move together sufficiently for the aggregate WSQS to separate lower- and higher-service-quality households.

Our supplementary measurement diagnostics further supported the underlying WSQS structure. The 32 WSQS-related survey items showed excellent sampling adequacy ($KMO = 0.937$), significant factorability (Bartlett's $\chi^2(496) = 40,065.6$, $p < 0.001$), and substantial overall item-set coherence (Cronbach's $\alpha = 0.850$). Parallel analysis identified six empirical factors, while no highly correlated item pairs ($|\rho| \geq 0.80$) or problematic cross-loadings were observed. Together with the close agreement across weighting architectures, these results indicate that the WSQS is both multidimensional and internally coherent, while avoiding excessive redundancy among its constituent items (Supplementary Fig. S1; Supplementary Tables S3–S4).

Distributional comparisons across weighting schemes provided the third structural check. The empirical cumulative distribution functions for equal-weight, population-priority, and individual-priority WSQS nearly overlapped (Fig. 3c), and pairwise agreement was high (Table 3). Equal-weight and population-priority WSQS were nearly identical, with $r = 1.000$ and mean absolute difference of 0.14 points. Individual-priority WSQS remained aligned with equal-weight WSQS, with $r = 0.991$ and mean absolute difference of 0.95 points. These results show that priority weighting can shift some household scores without destabilizing the aggregate WSQS distribution. Together, these structural checks justify using equal-weight WSQS as the primary non-circular outcome and treating priority-weighted WSQS versions as sensitivity outcomes.

WSQS domain structure and weighting robustness

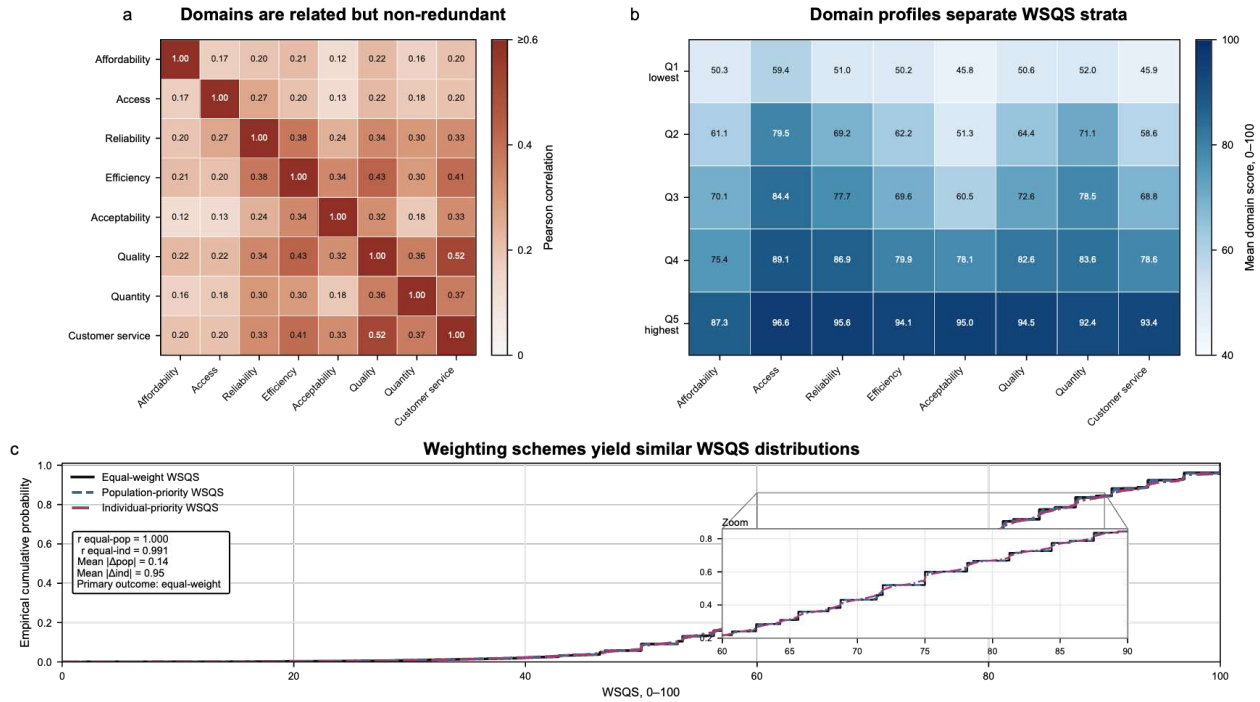


Figure 3. WSQS structure is coherent across domains and stable across weighting schemes.

a) Pearson correlations among the eight CORE-8 WSQS domain scores, with tile values shown. This panel tests whether the domains are related but non-redundant. b) Mean domain scores across equal-weight WSQS quintiles, with tile values shown. This panel tests whether the aggregate score coherently separates households across the underlying domains. c) Empirical cumulative distributions of equal-weight, population-priority-weighted, and individual-priority-weighted WSQS. This panel tests whether the aggregate distribution is sensitive to weighting architecture. WSQS scores are reported on the 0–100 scale, where higher values indicate better service quality.

Table 3. Robustness across WSQS weighting schemes.

Outcome	Mean	SD	Pearson with equal-weight	Mean absolute difference from equal-weight
Equal-weight WSQS	71.74	15.79	1.000	0.00
Population-priority WSQS	71.74	15.81	1.000	0.14
Individual-priority WSQS	71.77	15.93	0.991	0.95

Service-support variables explain most predictable variation in WSQS

Model performance was evaluated on the 20% test set, which was held out during model training. The mean-only baseline had a test-set RMSE of 0.6389 on the raw 1–5 severity scale and explained essentially no test-set variation ($R^2 \approx 0$) (Fig. 4a,b; Table 4). Adding service-support variables produced the largest improvement: the M1 service-support model reduced test-set RMSE to 0.4188 (Fig. 4a), increased test-set R^2 to 0.570 (Fig. 4b), and achieved a Spearman rank correlation of approximately 0.748.

Adding household-context variables produced little additional improvement. The M2 model, which included service-support and household context, had a test-set RMSE of 0.4187, R^2 of 0.571, and Spearman ρ of approximately 0.749 (Fig. 4a,b; Table 4). Adding stated-priority variables in the final M3 model produced a small additional gain: test-set RMSE decreased to 0.4179, R^2 increased to 0.572, and Spearman ρ remained approximately 0.749 (Fig. 4a,b; Table 4).

This incremental priority signal is not mechanically induced by outcome construction. The primary modeled outcome is the equal-weight WSQS, and stated priorities are not used to construct this primary outcome. Priorities enter only as predictors in M3 and as weights in separate sensitivity outcomes. Therefore, the M2-to-M3 comparison tests whether stated priorities explain residual variation in equal-weight WSQS beyond service-support and context variables.

The observed-versus-predicted plot on the 0–100 reporting scale showed that the final M3 model preserved the broad ordering of households by service quality, although prediction errors remained for individual households (Fig. 4c). The final M3 model had a reporting-scale RMSE of 10.45, R^2 of 0.572, and Spearman ρ of 0.749 (Fig. 4c). This nested model pattern supports two results. First, service-support conditions contain most of the predictable information in equal-weight WSQS. Second, household context and stated priorities add small but nonzero information beyond service-support variables, but the magnitude of these additions is modest relative to the initial gain from service-support predictors.

Table 4. Held-out model performance across nested predictor sets.

Model	Predictor blocks	Predictors	RMSE	MAE	R^2	Pearson r	Spearman ρ	RMSE 0–100
M0	Mean-only baseline	0	0.6389	0.522	0.000	–	–	15.97
M1	Service-support	25	0.4188	0.320	0.570	0.755	0.748	10.47
M2	Service-support + context	29	0.4187	0.319	0.571	0.755	0.749	10.47
M3	Service-support + context + priorities	37	0.4179	0.319	0.572	0.757	0.749	10.45

Nested prediction of equal-weight WSQS

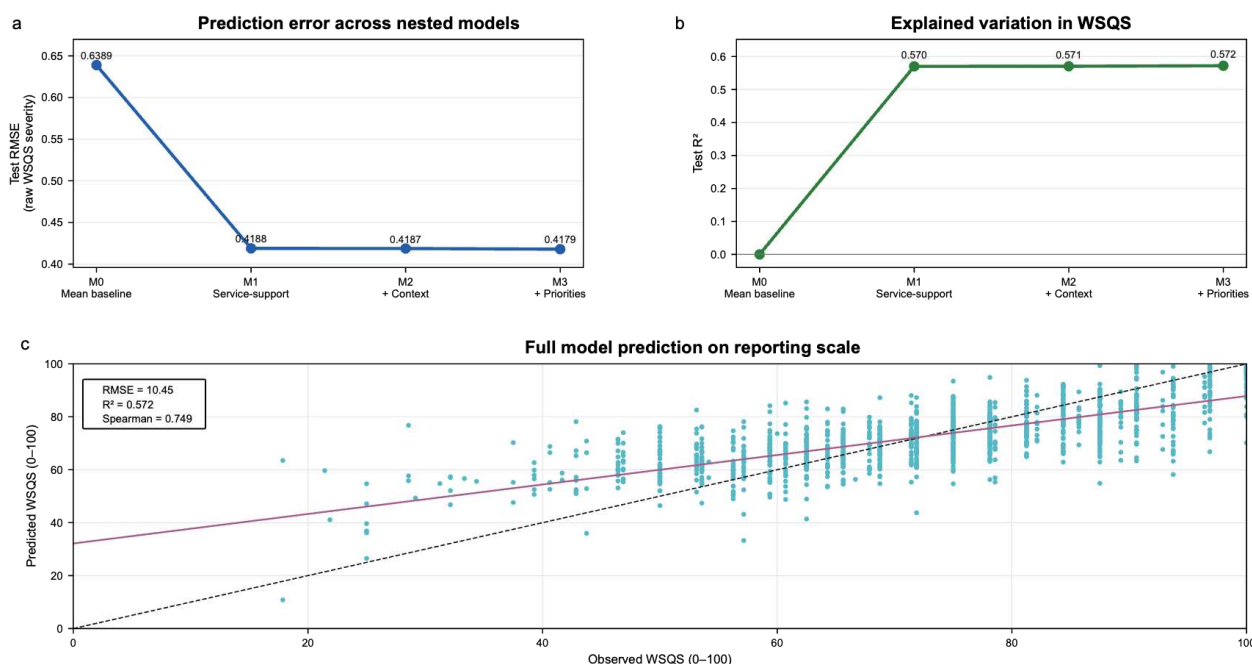


Figure 4. Service-support variables explain the most predictable variation in equal-weight WSQS.

a) Test RMSE across the nested model sequence, evaluated on the raw 1–5 WSQS severity scale. b) Test-set R² across the nested model sequence. c) Observed versus predicted equal-weight WSQS on the 0–100 reporting scale for the final M3 model. Context includes source/regime, ordered income, household size, and dwelling type; age and gender are evaluated only in sensitivity analyses.

Operational and institutional service conditions dominate model interpretation

The SHAP interpretation results confirmed that model predictions were dominated by operational and institutional service-support variables. In the primary 20% test set, service-support variables accounted for 94.9% of total grouped mean absolute SHAP attribution, compared with 4.1% for stated priorities and 0.9% for household context (Fig. 5a; Table 5). This pattern indicates that most prediction magnitude in the final M3 model came from reported service-support conditions rather than household background variables or stated priorities.

The highest-ranked individual predictors were also service-support variables. The leading predictors were value for money, complaint resolution, water pressure, appearance acceptability, water clarity, provider contact, shortage experience, quality change, backup reliability, taste acceptability, weekly water volume, and disruption notice (Fig. 5b; Table 6). These predictors span both physical and institutional dimensions of service: water pressure, clarity, quantity, shortages, and appearance describe direct service conditions, while complaint resolution, provider contact, value for money, backup reliability, and disruption notice capture responsiveness and service support.

This attribution pattern is important because it shows that WSQS variation was not explained primarily by household context or stated priorities. Instead, the most significant signals were the operational and institutional conditions through which households experienced water service. The grouped attribution result should be interpreted as block-level predictive attribution rather than as a claim that every service-support variable is individually more important than every context or priority variable. The robustness checks later on further evaluate whether this pattern is stable across repeated data partitions.

Predictive attribution and service-quality heterogeneity

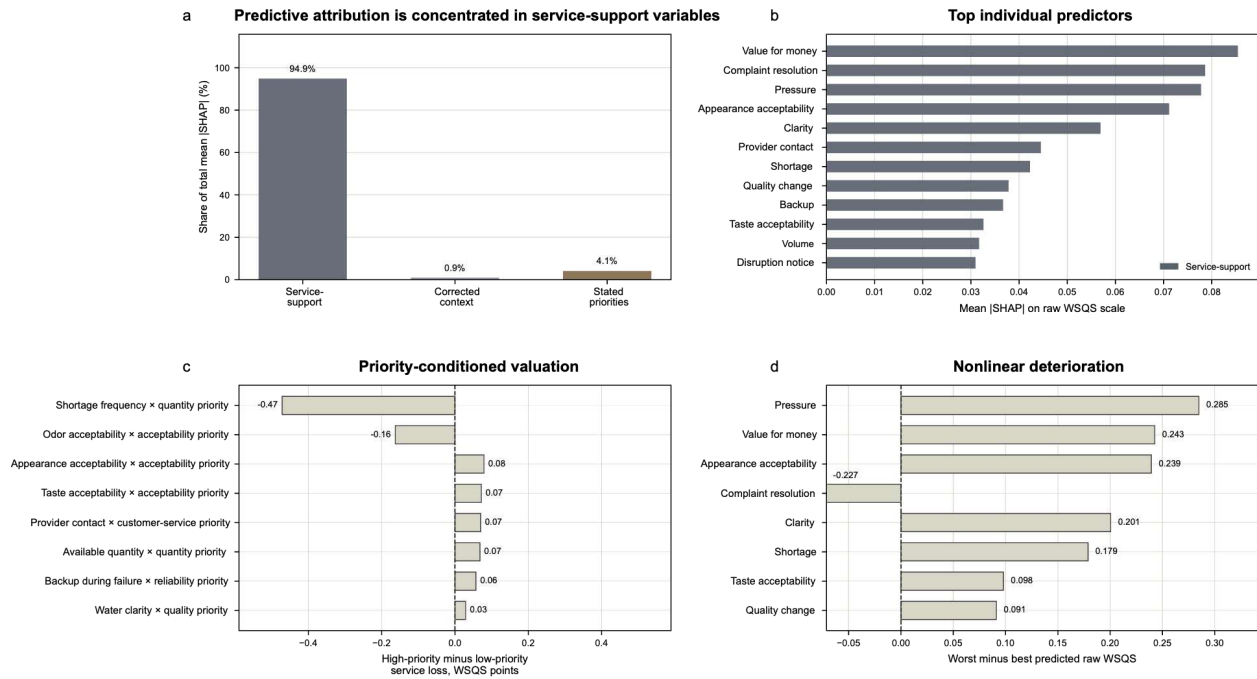


Figure 5. Predictive attribution and service-quality heterogeneity.

a) Grouped SHAP attribution in the final M3 model, showing the share of total mean absolute SHAP accounted for by service-support, household-context, and stated-priority variables. b) Top individual predictors by mean absolute SHAP on the raw WSQS scale. c) Priority-conditioned service-response contrasts, reported as the high-priority minus low-priority difference in predicted service loss on the 0–100 WSQS scale. Negative values indicate smaller predicted service loss among high-priority households, while positive values indicate larger predicted service loss. d) Nonlinear service-response contrasts reported as worst-minus-best predicted raw WSQS; larger positive values indicate higher predicted raw WSQS severity as the service condition worsens; negative values indicate non-monotonic endpoint behavior and should be interpreted from the full response curve.

Table 5. Grouped SHAP attribution in the final M3 model.

Predictor block	Feature count	Total mean absolute SHAP 0–100 equivalent	Share of total SHAP
Service-support	25	0.808	20.20
Context	4	0.008	0.20
Stated priorities	8	0.035	0.88

Note. Grouped SHAP shares are calculated by summing mean absolute SHAP values within each predictor block and dividing by the total mean absolute SHAP contribution across all predictors.

Table 6. Top predictors by mean absolute SHAP in the final M3 model.

Rank	Predictor	Predictor block	Mean absolute SHAP	0–100 equivalent	Share of total SHAP
------	-----------	-----------------	--------------------	------------------	---------------------

1	Value for money	Service-support	0.085	2.14	10.0%
2	Complaint resolution	Service-support	0.079	1.97	9.2%
3	Water pressure	Service-support	0.078	1.95	9.1%
4	Appearance acceptability	Service-support	0.071	1.78	8.4%
5	Water clarity	Service-support	0.057	1.42	6.7%
6	Provider contact	Service-support	0.045	1.11	5.2%
7	Shortage experience	Service-support	0.042	1.06	5.0%
8	Quality change	Service-support	0.038	0.95	4.4%
9	Backup reliability	Service-support	0.037	0.92	4.3%
10	Taste acceptability	Service-support	0.033	0.81	3.8%
11	Weekly water volume	Service-support	0.032	0.79	3.7%
12	Disruption notice	Service-support	0.031	0.77	3.6%
13	Supply schedule	Service-support	0.029	0.72	3.4%
14	Odor acceptability	Service-support	0.029	0.71	3.4%
15	Infrastructure maintenance	Service-support	0.027	0.68	3.2%

Note. Values are mean absolute SHAP values from the final M3 model using equal-weight CORE-8 WSQS as the primary outcome. Larger SHAP values indicate a greater contribution to the magnitude of the model's prediction, not causal importance.

Stated priorities add modest but interpretable heterogeneity

Although stated-priority variables contributed less than service-support variables overall, they provided a small nonzero signal in the final M3 model. In the primary 20% test set, stated priorities accounted for 4.1% of grouped SHAP attribution (Fig. 5a; Table 5). Because the primary outcome was equal-weight WSQS, these priority contributions are non-circular; stated priorities were not used to construct the dependent variable. Instead, they test whether households' stated valuations of service domains explain additional variation in equal-weight WSQS after observed service-support conditions and household context are included.

Priority-conditioned service-response contrasts clarified where priorities modestly shaped model predictions (Fig. 5c; Table 7). In the table and figure, the "x" symbol identifies the service condition and the corresponding stated-priority dimension used for the contrast. Each contrast compares the predicted service-response pattern for households with higher versus lower stated priority in the corresponding domain.

Using the 0–100 WSQS service-loss scale, the largest priority-conditioned difference was for shortage frequency × quantity priority: predicted service loss was 4.72 points among low-priority households and 4.25 points among high-priority households, yielding a high-minus-low difference of −0.47 points (Fig. 5c; Table 7). Odor acceptability × acceptability priority showed a smaller negative difference of −0.16 points. The remaining priority-conditioned differences were close to zero, ranging from 0.03 to 0.08 WSQS points for appearance acceptability, taste acceptability, provider contact, available quantity, backup during failure, and water clarity. These results indicate that stated priorities modestly condition selected service-response patterns, but the differences are small relative to the main effects of service-support conditions.

Detailed priority-conditioned curves show the same pattern (Fig. 6a–d). Quantity, affordability, odor, and appearance relationships differ slightly across low-, middle-, and high-priority groups, but the curves remain close together. Thus, priorities refine interpretation but do not drive the model. The dominant empirical signal remains the service condition itself, while priorities provide a smaller layer of heterogeneity in how service conditions align with households' stated importance. Extended predictor rankings, service-priority pair definitions, nonlinear response summaries, and the survey-question map are provided in Supplementary Section S3.

Priority-conditioned and nonlinear service-quality effects

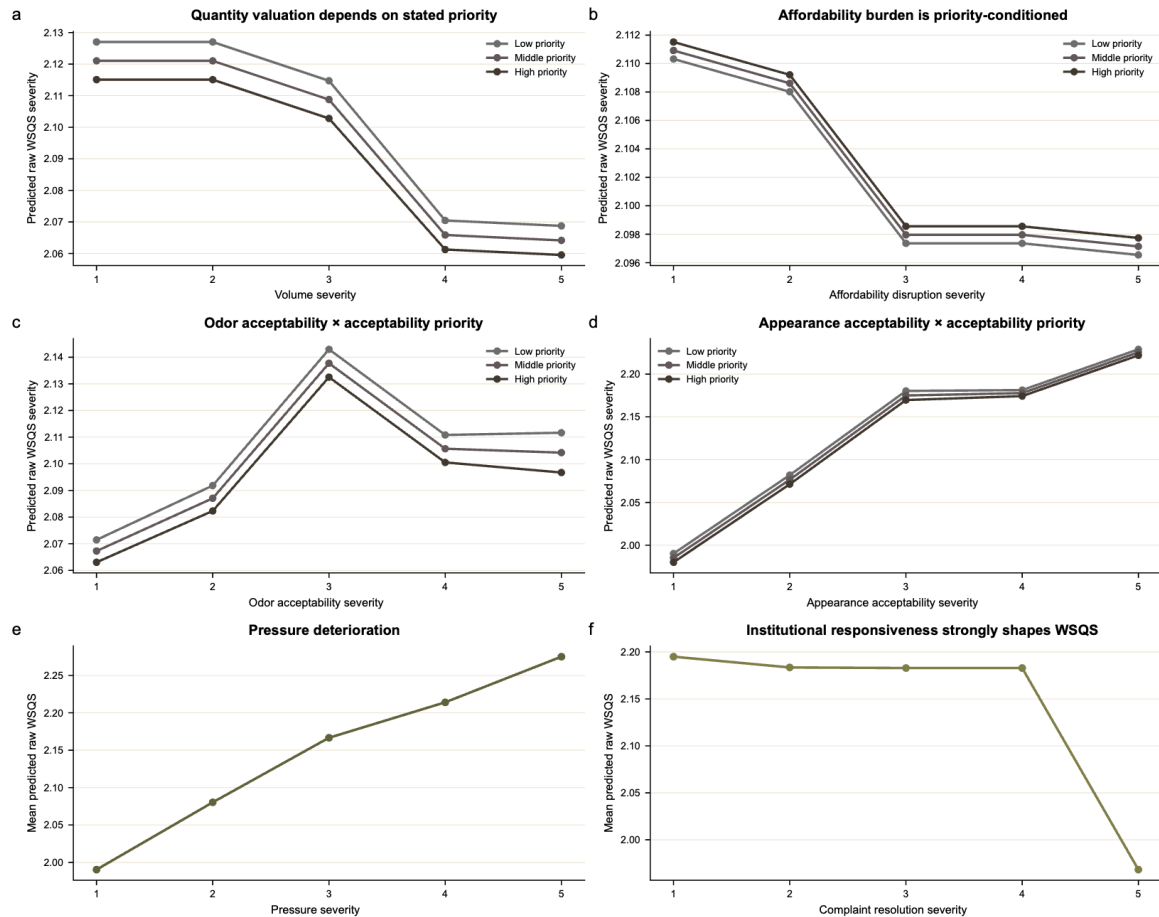


Figure 6. Priority-conditioned and nonlinear service-quality effects.

a) Predicted raw WSQS severity across available-quantity severity levels, stratified by quantity priority. b) Predicted raw WSQS severity across affordability-disruption severity levels, stratified by affordability priority. c) Odor acceptability by acceptability priority. d) Appearance acceptability by acceptability priority. e) Water pressure response curve showing increasing predicted raw WSQS severity as reported pressure conditions worsen. f) Complaint-resolution curve, showing a category-specific institutional responsiveness pattern. **Note.** All service-condition variables are shown on the harmonized 1–5 severity scale, where 1 indicates the best reported condition and 5 indicates the worst reported condition. Predicted raw WSQS uses the same direction: lower values indicate better predicted service quality and higher values indicate worse predicted service quality.

Table 7. Priority-conditioned service-response contrasts on the 0–100 WSQS scale.

Service-priority pair	Low-priority service loss, WSQS points	High-priority service loss, WSQS points	High-minus-low difference, WSQS points
-----------------------	--	---	--

Shortage frequency × quantity priority	4.72	4.25	−0.47
Odor acceptability × acceptability priority	1.01	0.84	−0.16
Appearance acceptability × acceptability priority	5.97	6.05	0.08
Taste acceptability × acceptability priority	2.40	2.47	0.07
Provider contact × customer- service priority	1.93	2.00	0.07
Available quantity × quantity priority	−1.46	−1.39	0.07
Backup during failure × reliability priority	2.02	2.08	0.06
Water clarity × quality priority	4.97	5.00	0.03

Note. Endpoint service loss is the decline in predicted WSQS on the 0–100 scale from the best to worst service-condition level. Positive values indicate lower predicted WSQS as the service condition worsens; negative values indicate non-monotonic or category-specific endpoint patterns and should be interpreted from the full response curve.

Key service-support conditions show nonlinear deterioration

The nonlinear response analyses showed that several service-support variables had nonlinear or category-specific service-response patterns with predicted WSQS (Fig. 5d; Fig. 6e,f). The largest deterioration occurred for water pressure: predicted raw WSQS severity increased by 0.285 from the best to worst severity level (Fig. 5d, 6e). On the 0–100 reporting scale, this corresponds to approximately a 7.13-point decline in predicted WSQS. The pressure curve showed steady deterioration, with a visibly sharper mid-range increase in predicted raw severity (Fig. 6e). Other large nonlinear contrasts included value for money (0.243 raw WSQS), appearance acceptability (0.239), water clarity (0.201), and shortage experience (0.179) (Fig. 5d). These results indicate that changes in household-reported service conditions were not simply linear across response categories.

Complaint resolution was among the highest-ranked SHAP predictors, but its nonlinear response curve was category-specific rather than simply monotonic (Fig. 5d; Fig. 6f). This pattern means the full response curve should be used for interpretation rather than relying only on the endpoint contrast. More generally, these nonlinear analyses are interpretive model summaries. They identify service conditions for which predicted WSQS changes unevenly across severity levels, but they do not establish causal patterns.

Together, the priority-conditioned and nonlinear results support the interpretation that WSQS is shaped primarily by observed service-support conditions, with modest priority-conditioned heterogeneity.

Robustness and sensitivity analyses support the model interpretation

The robustness analyses supported the stability of the main findings across data partitions, model classes, context encodings, attribution checks, and WSQS outcome definitions (Fig. 7). Repeated train-test splits

were used to assess whether the main model pattern depended on a single favorable random data partition. In this analysis, the dataset was repeatedly divided into new training and 20% test sets, and the nested model sequence was re-estimated on each split. Across these repeated splits, most of the predictive gain consistently occurred when service-support variables were added to the mean baseline: mean RMSE decreased from 0.6237 for M0 to 0.4150 for M1 (Fig. 7a). Adding context variables changed RMSE only slightly (0.4151), while adding stated priorities reduced RMSE to 0.4125 (Fig. 7a). Extended repeated-split, benchmark-model, context-specification, and attribution results are reported in Supplementary Section S2.

Household rank-ordering was also stable across repeated splits. Spearman rank correlation was 0.733 for M1, 0.732 for M2, and 0.735 for M3 (Fig. 7b). This indicates that the service-support model already captured most of the rank-order information in WSQS, while context and priority predictors produced only small changes.

Benchmark comparisons showed that CatBoost performed best or comparably under the final M3 predictor set. Across repeated splits, CatBoost had a mean RMSE of approximately 0.413, compared with approximately 0.424 for Histogram Gradient Boosting, 0.425 for Extra Trees, 0.427 for Random Forest, and 0.431 for Ridge and ElasticNet (Fig. 7c). These findings support using CatBoost as the primary learner while showing that the general performance pattern was not dependent on a single model class.

Outcome robustness checks showed similar performance across equal-weight, population-priority, and individual-priority WSQS outcomes when using the final M3 predictor set (Fig. 7d; Table 8). Equal-weight and population-priority outcomes had nearly identical repeated-split RMSE values of approximately 0.413, while the individual-priority outcome had an RMSE of approximately 0.421. These outcome checks are interpreted as sensitivity analyses, not as evidence that priority predictors have independent explanatory value when the outcome itself is priority-weighted. The non-circular test of priority contribution is the M2-to-M3 comparison using the primary equal-weight WSQS outcome.

Context-specification checks also showed that alternative encodings of household context did not materially change the model pattern (Fig. 7e). Compared with the service-only model, main-context, full-context, and income-categorical variants changed RMSE by approximately -0.0011 to 0.0005 , confirming that context specification did not drive the result. Finally, repeated-split grouped SHAP analyses preserved the same attribution structure as the primary split; service-support variables accounted for 93.2% of grouped attribution, stated priorities for 4.7%, and context for 2.0% (Fig. 7f).

Together, these results indicate that the main findings are not artifacts of one weighting scheme, train-test split, model class, context encoding, or attribution calculation. The evidence supports the central claim that end-user water service quality in heterogeneous provision systems is multidimensional, operationally structured, and modestly conditioned by household priorities.

Table 8. Robustness across alternative WSQS outcome definitions.

Outcome	Role	RMSE	MAE	R ²	Spearman ρ	Mean observed raw	Mean predicted raw	Mean observed 0–100	Mean predicted 0–100
Equal-weight CORE-8 WSQS	Primary non-circular outcome	0.413	0.317	0.561	0.738	2.093	2.100	72.69	72.49
Population- priority CORE-8 WSQS	Population- priority sensitivity outcome	0.413	0.317	0.563	0.741	2.092	2.101	72.69	72.49

Individual- priority CORE-8 WSQS	Household- priority sensitivity outcome	0.421	0.323	0.554	0.734	2.091	2.100	72.72	72.50
--	--	-------	-------	-------	-------	-------	-------	-------	-------

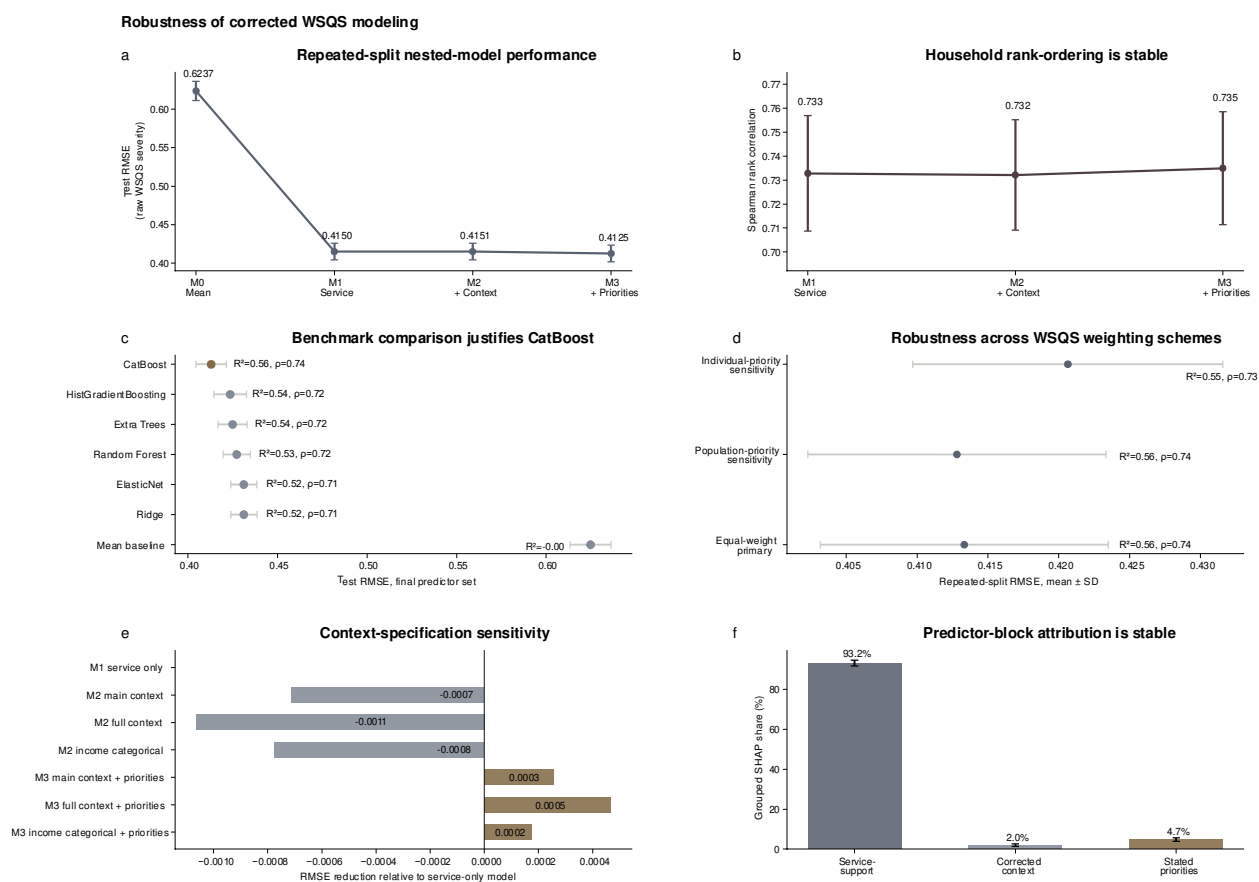


Figure 7. Robustness of WSQS modeling.

a) Repeated-split nested-model performance across the modeling sequence. Test RMSE on the raw 1–5 WSQS severity scale decreased sharply from the mean-only baseline to the service-support model, while adding context and stated priorities produced only small additional changes. b) Spearman rank correlation across repeated splits, showing that household rank-ordering was stable across the service-support, context, and priority-augmented models. c) Benchmark comparison across model classes using the final predictor set, showing that CatBoost performed best or comparably while the main predictive pattern was not dependent on a single learner. d) Robustness across WSQS weighting schemes, comparing equal-weight, population-priority, and individual-priority WSQS outcomes under repeated splits. e) Context-specification sensitivity, reported as RMSE reduction relative to the service-only model; alternative context encodings produced negligible changes in predictive performance. f) Repeated-split grouped SHAP attribution, showing that service-support variables consistently accounted for most predictive attribution, while context and stated priorities contributed smaller shares. **Note.** Error bars show variation across repeated train-test splits. RMSE is reported on the raw WSQS severity scale, where lower values indicate better predictive performance. Grouped SHAP shares represent the share of total mean absolute SHAP attribution within each predictor block.

Discussion

This study develops a household-level framework for measuring end-user water service quality in heterogeneous provision systems. The central finding is that household water service quality is multidimensional, structured, and diagnosable. As summarized in Table 9, WSQS varied across affordability, access, reliability, efficiency, acceptability, perceived quality, quantity sufficiency, and customer service, confirming that source type, infrastructure access, or a single satisfaction question cannot adequately represent household service quality. At the same time, individual-priority weighting shifted some household scores but produced only limited aggregate change relative to the equal-weight score, with a mean absolute difference of 0.95 points and a correlation of 0.991. For monitoring, this means that user priorities can enrich interpretation without replacing a comparable baseline metric. Equal-weight WSQS provides a stable reference point across households, while priority-weighted scores identify where household values shift service-quality interpretation, especially in mixed-provision settings involving piped supply, groundwater, tanker delivery, commercial water, and other sources.

A key contribution of the framework is that it separates measurement from preference sensitivity. Equal-weight WSQS was retained as the primary outcome because it provides a transparent, non-circular baseline with stated priorities not embedded in the dependent variable. Population-priority and individual-priority scores were instead treated as sensitivity outcomes. Their similar aggregate distributions indicate that WSQS is stable, while household-level shifts indicate that priorities still matter more to some respondents than to others. This allows WSQS to remain comparable across households while preserving preference-aware interpretation.

The primary empirical result is that WSQS is primarily structured by operational and institutional service-support conditions. Adding service-support variables produced the main gain in prediction, while household context and stated priorities added only small incremental improvements. SHAP attribution reinforced this pattern, with service-support variables accounting for 94.9% of grouped attribution in the primary test set and 93.2% across repeated splits. Thus, WSQS functions more as a service diagnostic than as a demographic profiling tool. Specifically, household service quality is explained less by who households are than by the service conditions they experience.

Next, the leading predictors broaden the scope of what counts as water service quality. Physical and experiential conditions such as pressure, clarity, appearance, quantity, shortages, and taste mattered, but so did institutional dimensions such as value for money, complaint resolution, provider contact, backup reliability, disruption notice, and infrastructure maintenance. This shift monitoring beyond narrow quantity- or quality-centered measures toward a more complete view of how households experience service delivery, consistent with recent studies (Stoler et al. 2023; Pérez-Foguet, 2023; Fukuda et al., 2019). Further, household context added limited independent signal once direct service-support conditions were measured. This does not mean context is irrelevant; source type, income, household size, and dwelling conditions may shape exposure to service constraints. Rather, the result suggests that these contextual differences or inequities may operate through the operational conditions the model captures directly, such as pressure, shortages, affordability burden, provider contact, and complaint resolution. Direct measurement of service functioning is therefore more informative for diagnosing household-level water service quality than relying on household characteristics or infrastructure categories alone as proxies.

Stated priorities added a modest but interpretable layer of heterogeneity. Because the primary modeled outcome was an equal-weighted WSQS, this priority signal was not mechanically induced by priority-weighted outcome construction. The findings suggest that priorities refined how selected service problems were interpreted, especially when the priority domain aligned with the constrained service dimension. We found no evidence that stated priorities replace operational service conditions as the main predictors of WSQS.

Finally, the nonlinear response results further show why flexible, interpretable modeling is useful for this problem. Water pressure, value for money, appearance acceptability, water clarity, and experience with shortages showed the clearest nonlinear deterioration patterns. These findings suggest that service

deterioration may not be experienced as a smooth linear decline. For monitoring and planning, such patterns matter because average or linear summaries can obscure the points at which worsening service conditions produce sharper losses in household service quality. Also, the robustness checks strengthen the main interpretation. As summarized in Table 9, the same pattern held across weighting schemes, repeated train-test splits, benchmark learners, context encodings, and grouped attribution checks. This suggests that the results are not artifacts of a single weighting choice, model class, data split, or attribution method.

Several limitations remain in this work. The analysis is cross-sectional and predictive, not causal. Because all survey-based outcome and predictor measures were reported by the same household respondent, associations may partly reflect shared reporting tendencies as well as underlying service conditions. WSQS relies on self-reported household service experience and stated priorities, which may reflect recall, expectations, or reporting differences. Perceived water quality was measured through household experience rather than laboratory testing, so WSQS should complement, not replace, chemical or microbiological safety monitoring. The survey used a non-probability, purposive maximum-variation recruitment strategy; results should therefore not be interpreted as population-weighted prevalence estimates for a bounded geography. External validation is needed before transferring the calibrated model or weights to other regions or provision contexts.

Overall, the study shows that end-user water service quality in heterogeneous provision systems is multidimensional, operationally structured, modestly priority-conditioned, and robust across modeling choices. For water-service monitoring in LMICs and heterogeneous provision systems, this shifts the question from whether households have access to water to how well that service functions from the user's perspective.

Table 9. Summary of main empirical findings.

Result theme	Key evidence	Interpretation
WSQS is multidimensional	Domain means ranged from 64.93 for acceptability to 76.08 for reliability	Service quality varies across domains and is not captured by source type, access, or satisfaction alone
Equal-weight WSQS is the primary outcome	Equal-weight, population-priority, and individual-priority WSQS means were 71.74, 71.74, and 71.77	Priority weighting changes some household scores but does not destabilize aggregate WSQS measurement
Service-support dominates prediction	M1 reduced test-set RMSE from 0.6389 to 0.4188; service-support SHAP share was 94.9% in the primary test set and 93.2% across repeated splits	Operational and institutional service conditions explain most predictable WSQS variation
Context adds limited incremental signal	Adding context changed test-set RMSE from 0.4188 to 0.4187; repeated-split context SHAP share was 2.0%	Household context contributes limited additional signal once direct service-support conditions are observed
Priorities add modest non-circular signal	Adding priorities improved test-set RMSE from 0.4187 to 0.4179; priority SHAP share	Stated priorities refine interpretation but do not replace

	was 4.1% in the primary test set and 4.7% across repeated splits	service-support conditions as the main drivers
Priority effects are domain-specific but modest	Priority-conditioned service-response contrasts were small, with the largest 0–100 high-minus-low service-loss difference equal to -0.47 WSQS points	Priorities modestly condition selected service relationships but do not drive the overall model
Some service effects are nonlinear	Water pressure, value for money, appearance acceptability, water clarity, and shortage experience showed the largest nonlinear deterioration contrasts	Key service conditions show nonlinear or category-specific response patterns rather than simple linear changes
Findings are robust	Results were stable across repeated splits, benchmark learners, alternative WSQS outcomes, and grouped SHAP attribution checks	The main findings are not artifacts of one weighting scheme, split, model class, or context specification

Methods

Study design and analytical workflow

This study developed a household-level measurement and predictive modeling framework for end-user water service quality in heterogeneous water provision systems. The objective was to test whether a multidimensional Water Service Quality Score (WSQS) could represent household service experience more comprehensively than source type, infrastructure access, or a single satisfaction measure.

The workflow included (i) survey quality assurance, (ii) response harmonization, (iii) construction of the CORE-8 WSQS, (iv) comparison of alternative weighting architectures, (v) predictive modeling with service-support, household-context, and stated-priority variables, and (vi) model interpretation through SHAP, nonlinear response analysis, and robustness checks. The analysis was designed as a predictive and diagnostic modeling exercise, not as a causal treatment-effect study. Estimated relationships are therefore interpreted as associations between observed household service conditions, stated priorities, household context, and WSQS.

Study setting, sampling strategy, survey administration, and consent

The study used household survey data collected primarily in Tamil Nadu, South India, where domestic water service is provided through heterogeneous arrangements including piped supply, groundwater, tanker delivery, commercially purchased water, and combinations of primary and supplementary sources. The household was the unit of analysis. Adult respondents (≥ 18 years) served as household informants, reporting on household water-service conditions, constraints, coping arrangements, and priorities. This use of an adult household informant is consistent with established household-survey practice, including major international household surveys that collect household-level water and sanitation information through an adult household member (UNICEF MICS, 2018). Primary household data were necessary because source classification and nominal infrastructure access alone do not jointly capture experienced dimensions of service such as affordability, sufficiency, acceptability, and responsiveness (Chen et al., 2017; Omarova et al., 2019).

The survey used a non-probability, purposive maximum-variation recruitment strategy to obtain broad coverage of contrasting water-service conditions and provision arrangements (Palinkas et al., 2015; Penrod et al., 2003; Platt et al., 2006). Recruitment was conducted by the study investigator through supervised, investigator-administered field access, with local institutional partners facilitating access to potential participants and study communities. Structured chain-referral recruitment extended participation beyond the initial access points. The resulting household sample included respondents reporting on a range of water-provision arrangements and socioeconomic conditions, including formal, self-supplied, purchased, and supplementary water systems. Recruitment through multiple supervised access pathways broadly followed the community-based household survey approaches described by Cooper et al. (2023), Banerji et al. (2013), and Hori et al. (2013). Local institutional partners facilitated field access and recruitment but did not define the analytical population.

This sampling design was selected to capture heterogeneity in household water-service experience rather than to estimate population prevalence. Because individual inclusion probabilities were unknown, the study does not make design-based population estimates for Tamil Nadu. Analyses therefore focus on measurement structure, heterogeneity in observed household service conditions, and predictive relationships across the water-service configurations represented in the dataset. Respondent characteristics are treated as descriptive attributes of the household informants rather than as the inferential target of the study. The Supplementary Information (Section S4) provides an independent sample-coverage and external-benchmark diagnostic, characterizing household composition, socioeconomic indicators, geographic reach and recruitment concentration, rural–urban settlement context, population-density coverage, and the diversity of domestic water-provision arrangements represented in the study. These diagnostics clarify the empirical breadth of the sample while transparently distinguishing observed coverage from probability-based representativeness.

Survey deployment was preceded by a pilot involving 266 participants approximately one week before full-scale field deployment. The pilot assessed question comprehension, response-option clarity, survey flow, accessibility, and administration procedures, and the findings were used to refine the finalized field instrument before deployment (Presser et al., 2004; Willis, 2005; Groves et al., 2009). Translated versions used during field administration were also reviewed and piloted before deployment. During the main study, surveys were investigator-administered using a standardized instrument and supervised field procedures.

Stanford University Institutional Review Board reviewed and approved the study protocol (protocol no. 82168) before data collection. All respondents were aged 18 years or older and provided electronic informed consent before participation. Participation was voluntary, and responses were de-identified before analysis.

Survey quality assurance and analytical sample

Survey records were screened before analysis using pre-specified quality-assurance rules. Records were retained only if they had sufficient completion, valid consent and eligibility information, interpretable household-level responses, and usable geographic information. Records were excluded if they exhibited very low completion rates, implausible durations, duplicate or bot-like patterns, synthetic or internally inconsistent responses, or other indicators of poor data quality (Supplementary Section S1.1). A high-completion threshold was applied so that retained records contained the core service-quality, priority, and context information needed for WSQS construction and modeling.

After quality assurance, cleaning, standardization, WSQS construction, and predictor processing, the final QA-passed analytical dataset contained 6,145 households. The primary modeling split used an 80:20 partition, with 4,916 households assigned to the training set and 1,229 to the held-out test set.

Survey measures

The survey captured four groups of variables. Specifically, first, CORE-8 service-quality items were used to construct WSQS across eight domains: affordability, access, reliability, efficiency, acceptability, perceived quality, quantity sufficiency, and customer service. One broad item from each domain was used to construct the primary WSQS. Second, service-support variables described operational and institutional service conditions not used directly in the CORE-8 outcome. These included value for money, water pressure, water clarity, acceptability of appearance, taste, and odor, quality change, backup during failure, shortage experience, available quantity, disruption notice, infrastructure maintenance, complaint resolution, provider contact, waiting time, supply schedule, cost, and related service experience variables. Third, respondent-priority variables measured the importance households assigned to each service domain on the original priority scale. These priorities were used as predictors in the final model and as weights in separate sensitivity outcomes, but not in the primary equal-weight WSQS outcome. And finally, household-context variables described household and housing conditions. The main context block included source/regime, ordered income, household size, and dwelling type. Age and gender were evaluated only in sensitivity analyses.

Response harmonization

Survey responses were harmonized before WSQS construction and model development. All CORE-8 service-quality items were mapped to a common 1–5 severity scale, where 1 represented the best service condition or lowest burden and 5 represented the worst service condition or highest burden. Items already ordered, from best to worst, retained their original order. Reverse-oriented items were recoded so that higher values consistently indicated worse service conditions. Responses outside valid survey categories, non-substantive responses, and structurally inapplicable responses were treated as missing for the relevant item. This harmonization preserved the ordinal survey structure while ensuring a consistent interpretation across domains. No attempt was made to convert ordinal responses into externally calibrated physical measurements; the scale represents household-reported service experience.

CORE-8 WSQS construction

The primary WSQS was constructed from one harmonized item in each of the eight service-quality domains. Let x_{id} denote the cleaned 1–5 severity-coded response for household i in domain d , where $d = 1, \dots, 8$. The equal-weight raw WSQS was calculated as:

$$WSQS_i^{raw,eq} = \frac{1}{8} \sum_{d=1}^8 x_{id}$$

The raw score is severity-oriented: lower values indicate better service quality, and higher values indicate worse service quality. For reporting and interpretation, the raw score was transformed to a 0–100 scale:

$$WSQS_i^{100} = 100 \times \frac{5 - WSQS_i^{raw}}{4}$$

Under this transformation, a raw score of 1 corresponds to 100 and a raw score of 5 corresponds to 0. Higher 0–100 values therefore indicate better service quality.

The equal-weight CORE-8 WSQS was retained as the primary modeling outcome because it provides a transparent and non-circular baseline. Stated priorities were not used to construct this primary outcome.

Priority-weighted sensitivity outcomes

Two alternative weighting architectures were constructed as sensitivity outcomes: population-priority-weighted WSQS and individual-priority-weighted WSQS.

For the population-priority outcome, the mean stated priority for each domain was calculated across households and normalized so that the eight domain weights summed to one:

$$w_d^{pop} = \frac{p_d^-}{\sum_{d'=1}^8 p_{d'}^-}$$

The population-priority raw WSQS was then calculated as:

$$WSQS_i^{raw,pop} = \sum_{d=1}^8 w_d^{pop} x_{id}$$

For the individual-priority outcome, each household's own stated priorities were normalized across the eight domains:

$$w_{id}^{ind} = \frac{p_{id}}{\sum_{d'=1}^8 p_{id'}}$$

The individual-priority raw WSQS was calculated as:

$$WSQS_i^{raw,ind} = \sum_{d=1}^8 w_{id}^{ind} x_{id}$$

These priority-weighted outcomes were used to test whether WSQS distributions and model performance were sensitive to weighting architecture. They were not used as the primary outcome because doing so would embed stated priorities directly into the dependent variable and make the priority-predictor analysis circular.

Measurement-structure diagnostics

As a supplementary diagnostic of the measurement framework, we evaluated cross-item similarity, redundancy, and latent structure among 32 ordinal survey items spanning the eight WSQS service domains. Harmonized responses were coded in a common direction, with lower values representing better service conditions and higher values representing greater service burden. Factorability was assessed using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity (Bartlett, 1954), and the number of empirical factors was evaluated using parallel analysis (Horn, 1965). We then applied minimum-residual exploratory factor analysis with oblimin rotation, allowing the empirical factors to correlate. These analyses were used only to evaluate measurement coherence and overlap; they were not used to redefine the eight conceptually specified WSQS domains, select CORE-8 items, or derive WSQS weights. As an additional coherence diagnostic of the primary construction, Cronbach's alpha was calculated across the eight CORE-8 indicators using complete observed responses, with the existing median-imputed measurement matrix used as a sensitivity analysis. These coefficients were interpreted descriptively as measures of cross-domain coherence rather than as evidence of a unidimensional reflective scale. Full implementation details and item-level diagnostics are reported in the Supplementary Information (S1.3).

Predictor groups and anti-circularity screening

The predictive analysis used three predictor blocks: service-support variables, household-context variables, and stated-priority variables. Let A_i denote service-support variables, C_i denote context variables, and P_i denote stated priorities. The full predictor set was:

$$X_i = (A_i, C_i, P_i)$$

An anti-circularity rule was applied before modeling. The eight CORE-8 items used to construct the primary equal-weight WSQS were excluded from all predictor sets. Service-support variables were retained only when they described related but distinct operational or institutional service conditions rather than direct duplicates of the outcome items. This allowed the model to test whether broader service-support conditions predicted equal-weight WSQS without mechanically reconstructing the dependent variable.

The main context block included source/regime, ordered income, household size, and dwelling type. Age and gender were excluded from the main context specification and evaluated only as sensitivity variables. The full CORE-8 item-to-predictor crosswalk and predictor-group architecture are provided in Supplementary Sections S1.2 and S2.1.

Predictive modeling

CatBoost regression was used as the primary learner because the dataset contained structured household survey variables, mixed numerical and categorical predictors, nonlinear response patterns, and possible interactions between service conditions and stated priorities (Prokhorenkova et al., 2018). Models were trained on the raw 1–5 WSQS severity scale.

The predictive task was:

$$Y_i = f(X_i) + \epsilon_i$$

where Y_i is the equal-weight raw WSQS for household i , X_i is the predictor vector, $f(\cdot)$ is the unknown prediction function, and ϵ_i is residual error.

The final model estimated:

$$\hat{Y}_i = \hat{f}(A_i, C_i, P_i)$$

Model outputs were interpreted as predictive associations, not causal effects.

Nested predictor-set analysis

A nested model sequence was used to evaluate the incremental information provided by service-support variables, context, and stated priorities:

$$M1: \hat{Y}_i = \hat{f}(A_i), M2: \hat{Y}_i = \hat{f}(A_i, C_i), M3: \hat{Y}_i = \hat{f}(A_i, C_i, P_i)$$

M1 included service-support variables only. M2 added household context. M3 added stated priorities. The M2-to-M3 comparison provides the non-circular test of whether stated priorities explain additional variation in equal-weight WSQS beyond service-support and context variables. A mean-only model, M0, was used as the baseline.

Model training and evaluation

The main model evaluation used the fixed 80:20 train-test split. All models were trained on the training set and evaluated on the held-out test set. Performance was measured using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), Pearson correlation, Spearman rank correlation, and the equivalent RMSE on the 0–100 reporting scale.

RMSE was calculated as:

649

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (Y_i - \hat{Y}_i)^2}$$

650

The coefficient of determination was calculated as:

651

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{test}} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N_{test}} (Y_i - \bar{Y}_{test})^2}$$

652

Spearman rank correlation was used to assess whether the model preserved the household ordering of observed service quality.

654

Benchmark models

655

CatBoost performance was compared with alternative learners using the same outcome, predictor set, training/test structure, and evaluation metrics. Benchmark learners included ridge regression, ElasticNet, random forest regression, extra trees regression, and histogram gradient boosting regression (Hoerl et al., 1970; Zou and Hastie, 2005; Breiman, 2001; Friedman, 2001). These comparisons tested whether the main performance pattern depended on the CatBoost model class.

660

Model interpretation

661

Model interpretation used SHAP-based attribution (Lundberg and Lee, 2017), grouped attribution, nonlinear response analysis, and priority-conditioned service-response contrasts.

663

SHAP values decomposed each prediction into predictor contributions:

664

$$\hat{Y}_i = \phi_0 + \sum_{j=1}^J \phi_{ij}$$

665

where ϕ_0 is the average prediction and ϕ_{ij} is the contribution of predictor j for household i . Global predictor importance was calculated as the mean absolute SHAP value:

667

$$I_j = \frac{1}{N} \sum_{i=1}^N |\phi_{ij}|$$

668

Grouped SHAP shares were calculated by summing mean absolute SHAP values within each predictor block and dividing by the total attribution across all predictors. These shares were used to assess whether predicted WSQS was primarily explained by service-support variables, context, or stated priorities.

671

Nonlinear response analyses examined whether predicted WSQS changed smoothly or sharply across observed severity levels of key service-support variables. For each selected service variable, model predictions were summarized across response levels while averaging over the observed distribution of other covariates. Deterioration was summarized as the difference in predicted raw WSQS between the worst and best observed service levels. Because the raw scale is severity-oriented, positive changes indicate worse predicted service quality.

677

Priority-conditioned service-response contrasts tested whether the predicted service loss associated with a given service condition differed between lower- and higher-priority households. For each service-priority pair, the predicted decline in 0–100 WSQS from the best to worst service-condition level was calculated separately for lower- and higher-priority strata. The reported contrast was:

681

$$Difference = Loss_{high\ priority} - Loss_{low\ priority}.$$

These contrasts were interpreted as model-based predictive heterogeneity rather than causal interaction effects.

Robustness and sensitivity analyses

Robustness checks evaluated whether the main conclusions were sensitive to data partitioning, model class, outcome definition, context specification, or attribution procedure. First, repeated train-test splits tested whether the nested model pattern depended on a single favorable partition. Second, benchmark models tested whether the results were specific to CatBoost. Third, the final M3 predictor set was evaluated across equal-weight, population-priority, and individual-priority WSQS outcomes. Fourth, context-specification checks compared the main context block with fuller context and alternative income encodings. Finally, repeated-split grouped SHAP analyses tested whether the service-support attribution pattern remained stable across data partitions.

Reporting convention and inference

Two WSQS scales were used. Model fitting, validation, SHAP attribution, and nonlinear response analyses were conducted on the raw 1–5 severity scale, where lower values indicate better service quality. Descriptive domain and aggregate WSQS summaries were reported on the transformed 0–100 scale, where higher values indicate better service quality. Priority-conditioned service-loss contrasts were also reported on the 0–100 WSQS scale to aid interpretation. All figures and tables identify whether values are reported on the raw severity scale or the 0–100 reporting scale.

All results should be interpreted as predictive and diagnostic. The sampling strategy supports analysis of measurement structure and service-quality relationships within a heterogeneous household dataset, but it does not support population-weighted prevalence claims for Tamil Nadu or any bounded administrative unit. External validation is required before transferring calibrated weights, predictor effects, or model performance estimates to other regions or provision contexts.

Data Availability

All data supporting the findings of this study are included in the paper and its Supplementary Information. Because the study uses human-subject survey data, the full de-identified dataset will be made available from the corresponding author upon reasonable request, subject to applicable ethical and data-protection requirements. A public de-identified version of the dataset will also be deposited in an accessible data repository upon publication, and the repository link will be provided in the published article. Direct identifiers and fine-scale geographic variables that could enable participant or household re-identification will not be released.

Funding

A.M. acknowledges support for this research from the Stanford King Center on Global Development through the Graduate Student Research Grant. The funders had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Author contributions

A.M. conceptualized the study, developed the methodology, coordinated data collection, curated and analyzed the data, produced the visualizations, and drafted the manuscript. K.K.O. contributed to study conceptualization and methodology, supervised the research, provided guidance on study design and analytical strategy, contributed to interpretation of the findings, and revised the manuscript. Both authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgment

The authors thank the Stanford King Center on Global Development for supporting the study's fieldwork. They also thank Dr. B. S. Murugan and Dr. Saravanan M. of M. Kumarasamy College of Engineering, Karur, Tamil Nadu, India; Dr. J. Premalatha and Dr. N. Surumbarkuzhali of Government Arts College, Salem, Tamil Nadu, India; and Dr. T. K. Kannan, Dr. Sitheswaran Nainamalai and Dr. B. Sanjay Gandhi of Gnanamani College of Technology, Namakkal, Tamil Nadu, India, for serving as local institutional partners and supporting the study. The authors also thank all participants who took part in the survey and contributed to this study and to the broader effort to quantify household water-service quality.

References

1. Banerji, R., Bhattacharjea, S. & Wadhwa, W. The Annual Status of Education Report (ASER). *Res. Comp. Int. Educ.* **8**, 387–396 (2013).
2. Bartlett, M. S. A note on the multiplying factors for various χ^2 approximations. *J. R. Stat. Soc. B* **16**, 296–298 (1954).
3. Bazaanah, P., Buthelezi, S. J. & Oppong, D. A. K. Qualitative study of drinking water, sanitation and hygiene access: perspectives from the Central Gonja District, Ghana, and Mtubatuba Municipality, South Africa. *J. Water Sanit. Hyg. Dev.* **14**, 1043–1065 (2024).
4. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
5. Charles, K. J., Nowicki, S. & Bartram, J. K. A framework for monitoring the safety of water services: from measurements to security. *npj Clean Water* **3**, 36 (2020).
6. Chen, Q. et al. Approaches to improving survey-weighted estimates. *Stat. Sci.* **32**, 227–248 (2017).
7. Cooper, B., Crase, L., Burton, M. & Pham, H. D. Survey data on preferences and attitudes towards participatory irrigation management in India and Pakistan. *Sci. Data* **10**, 152 (2023).
8. Daly, S. W. et al. Multiple water source use in low- and middle-income countries: a systematic review. *J. Water Health* **19**, 370–392 (2021).

9. Denantes, J. & Donoso, G. Factors influencing customer satisfaction with water service quality in Chile. *Utilities Policy* **73**, 101295 (2021).
10. Elliott, M. et al. Addressing how multiple household water sources and uses build water resilience and support sustainable development. *npj Clean Water* **2**, 6 (2019).
11. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
12. Fukuda, S., Noda, K. & Oki, T. How global targets on drinking water were developed and achieved. *Nat. Sustain.* **2**, 429–434 (2019).
13. Geleijnse, J., Rutten, M., de Villiers, D., Bamwenda, J. T. & Abraham, E. Enhancing water access monitoring through mapping multi-source usage and disaggregated geographic inequalities with machine learning and surveys. *Sci. Rep.* **13**, 13433 (2023).
14. Grafton, R. Q. et al. Rethinking responses to the world's water crises. *Nat. Sustain.* **8**, 11–21 (2025).
15. Greenwood, E. E. et al. Mapping safe drinking water use in low- and middle-income countries. *Science* **385**, 784–790 (2024).
16. Groves, R. M., Fowler, F. J. Jr, Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. *Survey Methodology*. 2nd edn. Wiley, Hoboken, NJ (2009).
17. Hamilton, A. L. et al. Resilient water infrastructure partnerships in institutionally complex systems face challenging supply and financial risk trade-offs. *Nat. Commun.* **15**, 7354 (2024).
18. Hannah, D. M. et al. Water and sanitation for all in a pandemic. *Nat. Sustain.* **3**, 773–775 (2020).
19. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
20. Hopkins, E. G. & Sowby, R. B. A qualitative definition of reliable water supply for public water systems. *Water* **16**, 3359 (2024).
21. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
22. Hori, S., Kondo, K., Nogata, D. & Ben, H. The determinants of household energy-saving behavior: survey and comparison in five major Asian cities. *Energy Policy* **52**, 354–362 (2013).
23. Jeandron, A. et al. Predicting quality and quantity of water used by urban households based on tap water service. *npj Clean Water* **2**, 23 (2019).
24. Jepson, W. E., Wutich, A., Collins, S. M., Boateng, G. O. & Young, S. L. Progress in household water insecurity metrics: a cross-disciplinary approach. *WIREs Water* **4**, e1214 (2017).
25. Kayser, G. L., Moriarty, P., Fonseca, C. & Bartram, J. Domestic water service delivery indicators and frameworks for monitoring, evaluation, policy and planning: a review. *Int. J. Environ. Res. Public Health* **10**, 4812–4835 (2013).
26. Kumar, T., Post, A. E., Ray, I., Otsuka, M. & Pardo-Bosch, F. From public service access to service quality: the distributive politics of piped water in Bangalore. *World Dev.* **151**, 105736 (2022).
27. Lee, G. O. et al. Household coping strategies associated with unreliable water supplies and diarrhoea in Ecuador, an upper-middle-income country. *Water Res.* **170**, 115269 (2020).
28. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).
29. Mohanavelu, A. & Osman, K. K. Ripples of injustice: unravelling water equity challenges on a global scale. *Environ. Res. Lett.* **19**, 081007 (2024).
30. Mohanavelu, A., Osman, K. K., Kothari, C. & France-Mensah, J. Quantitatively incorporating social equity in water network maintenance and rehabilitation decision-making. *J. Water Resour. Plan. Manag.* **150**, 04024039 (2024).
31. Nowicki, S., Koehler, J. K. L. & Charles, K. J. Including water quality monitoring in rural water services: why safe water requires challenging the quantity versus quality dichotomy. *npj Clean Water* **3**, 14 (2020).

32. Ogata, R., Matsumoto, S., Otsuka, T., Hattori, Y. & Kitawaki, H. Influence of customer satisfaction and perceived water service quality on urban water supply management: a systematic review. *Cleaner Responsible Consum.* **21**, 100389 (2026).
33. Omarova, A., Tussupova, K., Hjorth, P., Kalishev, M. & Dosmagambetova, R. Water supply challenges in rural areas: a case study from Central Kazakhstan. *Int. J. Environ. Res. Public Health* **16**, 688 (2019).
34. Osman, K. K. & Faust, K. M. Toward operationalizing equity in water infrastructure services: developing a definition of water equity. *ACS ES&T Water* **1**, 1849–1858 (2021).
35. Palinkas, L. A. et al. Purposeful sampling for qualitative data collection and analysis in mixed-method implementation research. *Adm. Policy Ment. Health* **42**, 533–544 (2015).
36. Penrod, J., Preston, D. B., Cain, R. E. & Starks, M. T. A discussion of chain referral as a method of sampling hard-to-reach populations. *J. Transcult. Nurs.* **14**, 100–107 (2003).
37. Pérez-Foguet, A. Broadening the water affordability approach to monitor the human right to water. *Cities* **143**, 104573 (2023).
38. Platt, L. et al. Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injecting drug users across nine studies in Russia and Estonia. *J. Urban Health* **83**, 39–53 (2006).
39. Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J. & Singer, E. Methods for testing and evaluating survey questions. *Public Opin. Q.* **68**, 109–130 (2004).
40. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **31** (2018).
41. Rajapakshe, P. S. K., Termansen, M. & Paavola, J. Public preferences for the improvement of existing piped water provision: a choice experiment in Sri Lanka. *Front. Water* **5**, 1134660 (2023).
42. Savelli, E., Mazzoleni, M., Di Baldassarre, G., Cloke, H. & Rusca, M. Urban water crises driven by elites' unsustainable consumption. *Nat. Sustain.* **6**, 929–940 (2023).
43. Skerker, J. et al. Urban water affordability crisis exacerbated by climate change. *Nat. Sustain.* (2026).
44. Stoler, J., Jepson, W. E., Brewis, A. & Wutich, A. Frontiers of household water insecurity metrics: severity, adaptation and resilience. *BMJ Glob. Health* **8**, e011756 (2023).
45. United Nations Children's Fund (UNICEF). *Multiple Indicator Cluster Surveys (MICS6): Instructions for Interviewers*. UNICEF, New York (2018).
46. Vásquez, W. F., Raheem, N., Quiroga, D. & Ochoa-Herrera, V. Household preferences for improved water services in the Galápagos Islands. *Water Resour. Econ.* **34**, 100180 (2021).
47. Vilarinho, H., Pereira, M. A., D'Inverno, G., Nóvoa, H. & Camanho, A. S. Water Utility Service Quality Index: a customer-centred approach for assessing the quality of service in the water sector. *Socioecon. Plann. Sci.* **92**, 101797 (2024).
48. Willis, G. B. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. SAGE Publications, Thousand Oaks, CA (2005).
49. Wutich, A. et al. Advancing methods for research on household water insecurity: studying entitlements and capabilities, socio-cultural dynamics, and political processes, institutions and governance. *Water Secur.* **2**, 1–10 (2017).
50. Young, S. L. et al. The Household Water InSecurity Experiences Scale: development and validation of a household water insecurity measure for low-income and middle-income countries. *BMJ Glob. Health* **4**, e001750 (2019).
51. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MohanaveluOsmanNPJCleanWaterFinalSupplementary.docx](#)