Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 May 2026
Revised 18 June 2026
Accepted 08 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 603–607

KEYWORDS:
Big data analytics, water resource management, disease-endemic regions, machine learning; remote sensing, Internet of Things, public health informatics.

Big Data Analytics for Optimizing Sustainable Water Resource Management in Disease-Endemic Regions: A Narrative Review

M. Buvana[1], Mary Rajam Vandana.V [2], V. Meera [3], P.Subhashini[4], Muthumayil K.[5], G Vinoth Chakkaravarthy[6]

[1] Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India.
[2] Department of Artificial Intelligence and Data Science, Parul Institute of Engineering and Technology, Faculty of Engineering and Technology, Parul University, Vadodara, Gujarat, India..
[3]Department of Master of Computer Applications, S.A. Engineering College ,Chennai,Tamil Nadu, India.
[4]Department of Artificial Intelligence and Machine Learning, Panimalar Engineering College, Poonamalle, Chennai -600123. India.
[5] Department of Computer Science and Engineering, Vel Tech Dr. Rangarajan R&D Institute of Science and Technology, Chennai, India;
[6] Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu. India.

ABSTRACT: In cholera-, typhoid fever-, and schistosomiasis-endemic regions, unsafe and irregularly monitored water supplies remain a major driver of waterborne and water-related disease. As of 2024, an estimated 2.1 billion people worldwide still lacked access to safely managed drinking water (WHO/UNICEF Joint Monitoring Programme, 2025). Conventional water-quality surveillance, based on scheduled grab-sampling and laboratory analysis, is often too slow to support timely public health action. This narrative review synthesizes recent literature on the application of big data analytics — Internet of Things (IoT) sensor networks, satellite remote sensing, and machine learning — to water resource management in disease-endemic settings. The reviewed literature indicates that machine learning models — particularly ensemble methods such as random forest and XGBoost — have reported high accuracy for forecasting cholera outbreaks and predicting water-quality parameters, while low-cost IoT sensors and satellite-derived climate variables are increasingly used to fill monitoring gaps in resource-limited settings. Recurrent limitations include a predominance of geographically narrow studies, limited integration of real-time multi-sensor data with clinical surveillance, and persistent infrastructure, connectivity, and data-governance barriers in the lowest-resource settings that carry the greatest disease burden. We conclude that big data analytics holds considerable promise for shifting water-related disease management from reactive to anticipatory, but that this promise remains only partially realized and depends on sustained investment in local infrastructure and data governance.

1. INTRODUCTION

Access to safe water remains inseparable from public health outcomes. Despite measurable progress over the past decade, the WHO/UNICEF Joint Monitoring Programme (JMP) reports that global coverage of safely managed drinking water rose from 68% to 74% between 2015 and 2024, yet 2.1 billion people — including 106 million who draw water directly from untreated surface sources — still lacked safely managed drinking water in 2024, with the largest gaps concentrated in low-income, rural, and fragile settings (WHO/UNICEF JMP, 2025).

These same settings frequently overlap with regions endemic to waterborne and water-related diseases such as cholera, typhoid fever, and schistosomiasis, where the causative pathogens or their intermediate hosts persist in poorly managed water systems (Campbell et al., 2020; Xue et al., 2023).

Conventional water-quality monitoring, which typically relies on scheduled grab-sampling followed by laboratory analysis, can take days to return results — a delay that may separate a low-cost preventive intervention from an outbreak requiring emergency response. Over the past decade, three technological developments have converged to offer an alternative: low-cost Internet of Things (IoT) sensors capable of continuous in-situ water-quality monitoring (de Camargo et al., 2023); satellite remote sensing providing standardized, repeated environmental data at broad spatial and temporal scales, often at no cost to the user (Teitelbaum et al., 2024); and machine learning methods capable of learning complex, nonlinear relationships between environmental drivers and both water-quality parameters and disease incidence (Yan et al., 2024; Campbell et al., 2020).

This review synthesizes recent literature at the intersection of these three developments, with a specific focus on applications relevant to disease-endemic regions. Because water is itself a natural resource whose management directly determines human health outcomes in these settings, the topic sits within the interdisciplinary, health-oriented scope of this journal, extending its coverage of natural-resource science into the domain of environmental data analytics and public health informatics. Our objective is to characterize (i) how big data and machine learning methods have been applied to water-quality prediction and monitoring; (ii) how satellite- and sensor-derived environmental data have been used to forecast water-related disease risk, with cholera as the most extensively studied case; and (iii) what infrastructural, methodological, and governance gaps remain before such systems can be reliably deployed in the resource-limited settings that carry the greatest disease burden.

2. REVIEW METHODOLOGY

2.1 Search Strategy and Sources

This narrative review draws on peer-reviewed journal articles, systematic reviews, and institutional reports (World Health Organization, UNICEF) identified through focused searches on machine learning applications in water-quality prediction, low-cost IoT water-quality sensing, satellite remote sensing for infectious disease ecology, and machine learning-based prediction of cholera and related waterborne disease outbreaks. Priority was given to contemporary (2019–2025) peer-reviewed sources, supplemented where necessary by earlier foundational work on remote sensing applications in disease surveillance. Searches were not restricted to a specific database or fixed time window.

2.2 Scope and Limitations of the Review Approach

As a narrative rather than systematic review, this paper does not apply formal PRISMA screening criteria, and the source set, while illustrative of major themes in the field, is not exhaustive. Reported quantitative results (e.g., model accuracy metrics) are drawn directly from the cited primary studies and are presented here as reported by their original authors; they have not been independently re-verified or re-analyzed by the present authors.

2.3 Conceptual Framework and System Architecture

Figure 1 summarizes the conceptual architecture underlying the reviewed literature. Data acquisition draws on three complementary streams — in-situ IoT water-quality sensors, satellite-derived environmental variables, and clinical/epidemiological surveillance records — which are harmonized and fed into machine learning models (most commonly random forest, XGBoost, and LSTM-based architectures) for water-quality prediction and disease-outbreak forecasting. Outputs are intended to support early-warning dashboards and, ultimately, decision-making in resource-limited, disease-endemic settings. This framework is a synthesis of the architectures reported across the reviewed studies rather than a single deployed system, and is offered here to orient the results discussed in Section 3.

Natural Resources for
Human Health
eISSN: 2583-1194

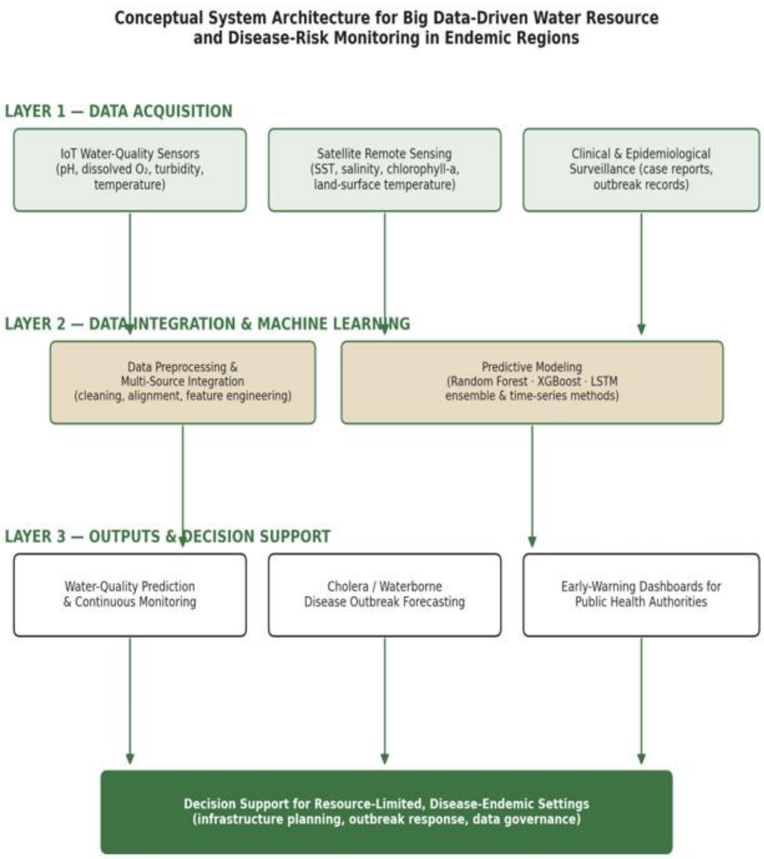

Figure 1. Conceptual system architecture synthesized from the reviewed literature, showing the flow from multi-source data acquisition through machine learning-based analytics to public health decision support.

3. RESULTS AND DISCUSSION

3.1 Low-Cost Sensing and IoT Infrastructure for Water Monitoring

A systematic review of low-cost water-quality sensors for IoT applications examined the sensor types, manufacturers, and measured parameters reported across the literature and found that the field is served by a small number of primary sensor vendors, with pH, dissolved oxygen, turbidity, and temperature the most commonly monitored parameters (de Camargo et al., 2023). Related work has demonstrated low-cost, real-time telemetry architectures — combining microcontrollers, water-quality sensors, and wireless connectivity — piloted in resource-limited settings such as El Salvador and South Africa, where conventional monitoring instrumentation is scarce or prohibitively expensive. In one South African deployment, a low-cost sonar-based IoT flow sensor installed downstream of an informal settlement revealed previously unknown drainage and peak-flow dynamics relevant to water planning in a setting disproportionately exposed to rainfall-driven contamination risk. Collectively, this literature indicates that low-cost IoT sensing is technically feasible in low-resource settings, though the reviewed systems remain largely at pilot or demonstration scale rather than sustained operational deployment.

3.2 Machine Learning for Water Quality Prediction

Owing to the shortcomings of numerical models in parameter selection, adaptability, and computational efficiency, machine learning has gradually replaced traditional numerical models for water-quality prediction, according to a thorough analysis of over 170 studies published over a five-year period (Yan et al., 2024). Recurrent and ensemble architectures, such as Long Short-Term Memory (LSTM) networks and tree-based ensembles, are commonly reported as the best-performing model classes for time-series water-quality data. The reviewed applications cover both single-indicator prediction (e.g., dissolved oxygen, turbidity) and composite water-quality-index prediction. These results are consistent with a separate review that evaluated 45 different machine learning algorithms for water-quality monitoring applications, underscoring the methodological diversity in this field and the absence of a single approach that is consistently dominant across contexts.

3.3 Satellite Remote Sensing for Environmental and Disease-Relevant Data

Remote sensing has an established role in infectious disease ecology. A quantitative review of the primary literature found that satellite-derived data are most commonly used to study geographically widespread human diseases, and that early warning systems for cholera based on remotely sensed sea surface temperature and salinity have allowed public health authorities to issue warnings before the first clinical case is detected (Teitelbaum et al., 2024). Complementary reviews of remote sensing applications in vector-borne and parasitic disease mapping describe extensive use of satellite-derived vegetation indices, land surface temperature, and land-cover change data in mapping malaria and schistosomiasis risk, and note growing use of low-cost drones for higher-resolution, near-real-time environmental data collection in outbreak-prone areas. More recent work has applied deep learning directly to high-resolution remote sensing imagery to identify and monitor environmental sources of schistosomiasis transmission, extending remote sensing's role from broad environmental risk mapping toward finer-grained, automated identification of transmission foci (Xue et al., 2023).

3.4 Machine Learning-Based Prediction of Waterborne Disease Outbreaks: The Cholera Case

Cholera has emerged as the most extensively studied case of machine learning applied to water-related disease prediction, reflecting both its strong environmental determinants and its continued public health burden. A random forest classifier trained on satellite-derived essential climate variables (land surface temperature, sea surface salinity, chlorophyll-a concentration, and sea level anomaly) against a 2010–2018 cholera outbreak dataset for coastal India reported an accuracy of 0.99, an F1 score of 0.942, and a sensitivity of 0.895, correctly identifying 89.5% of outbreaks in the test data (Campbell et al., 2020). A separate study developing a Cholera Outbreak Risk Prediction (CORP) model for Nigeria, combining dimensionality reduction, synthetic oversampling for class imbalance, and an XGBoost classifier, reported an accuracy of 99.62%, a Matthews correlation coefficient of 0.976, and an area under the curve of 99.2%, outperforming several previously published cholera prediction models compared within the same study (Amshi et al., 2024). Earlier work using an XGBoost-based ensemble (the Cholera Artificial Learning Model, or CALM) to forecast cholera incidence in Yemen over 2- to 8-week horizons reported forecast errors within approximately five cases per 10,000 people, using rainfall, historical incidence, and conflict-related covariates (Badkundri et al., 2019, as reported in subsequent reviews). Taken as a whole, these studies show a consistent pattern: models that include climatic or environmental covariates in addition to historical case data report high discriminative performance; however, comparability across studies is constrained by differences in validation strategy, geography, and outcome definitions.

3.5 Synthesis: Table of Reported Model Performance

Performance metrics from the key studies discussed above are compiled in Table 1. These figures were not independently validated; they are reproduced from the original publications for comparative reference.

Table 1. Reported performance of machine learning models in selected water-quality and cholera prediction studies (metrics as reported by the original authors).

Study	Disease/Domain	Model	Reported Performance
Campbell et al. (2020)	Cholera risk, coastal India	Random Forest classifier	Accuracy 0.99; F1 0.942; Sensitivity 0.895
Amshi et al. (2024)	Cholera outbreak risk, Nigeria	XGBoost (CORP model)	Accuracy 99.62%; MCC 0.976; AUC 99.2%
Badkundri et al. (2019)	Cholera incidence, Yemen	XGBoost ensemble (CALM)	Forecast error ≈ 5 cases/10,000 (2–8 week horizon)
de Camargo et al. (2023)	Water-quality sensing (review)	Systematic review of IoT sensors	3 primary low-cost sensor vendors identified across reviewed literature

3.6 Integration Challenges and Research Gaps

Despite promising reported performance, the reviewed literature identifies several unresolved gaps that are pertinent to disease-endemic, resource-limited settings: (1) most water-quality machine learning studies are geographically limited and trained on region-specific datasets, which constrains generalizability; (2) few studies directly integrate real-time, multi-sensor IoT data with clinical or epidemiological surveillance streams — cholera prediction models to date have relied mostly on retrospective case-report data paired with climate or satellite covariates rather than continuous water-quality sensor networks; and (3) deployment in the lowest-resource settings remains intermittent even where pilot systems have proven technically feasible (de Camargo et al., 2023).

4. CONCLUSIONS

The reviewed evidence indicates that big data analytics — spanning low-cost IoT sensing, satellite remote sensing, and machine learning — has demonstrated strong reported predictive performance for both water-quality parameters and water-related disease outbreaks, with cholera forecasting the most mature application area to date. Realizing this potential at scale in disease-endemic, resource-limited regions will require closing three gaps identified in this review: geographic generalizability of existing models, tighter integration of real-time sensor data with clinical surveillance, and sustained investment in infrastructure and data governance suited to local operating conditions. Future primary research combining continuous water-quality sensing with prospective outbreak surveillance, evaluated across multiple regions, would meaningfully extend the retrospective, single-region studies that currently dominate this literature.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

This article is a narrative literature review based on secondary data from previously published, publicly available studies and institutional reports. It did not involve human participants, animal subjects, or the collection of primary data; institutional ethical approval was therefore not required.

DATA AVAILABILITY STATEMENT

No new data were generated or analyzed in the preparation of this review. All data discussed are available in the cited publications.

REFERENCES

[1] Amshi, H. A., Prasad, R., Sharma, B. K., Yusuf, S. I., & Sani, Z. (2024). How can machine learning predict cholera: insights from experiments and design science for action research. *Journal of Water and Health*, 22(1), 21–35. https://doi.org/10.2166/wh.2023.026

[2] Badkundri, R., Valbuena, V., Pinnamareddy, S., Cantrell, B., & Standeven, J. (2019). Forecasting the 2017–2018 Yemen Cholera Outbreak with Machine Learning. *Preprint, arXiv:1902.06739*.

[3] Campbell, A. M., Racault, M-F., Goult, S., & Laurenson, A. (2020). Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables. *International Journal of Environmental Research and Public Health*, 17(24), 9378. https://doi.org/10.3390/ijerph17249378

[4] de Camargo, E. T., Spanhol, F. A., Slongo, J. S., da Silva, M. V. R., Pazinato, J., de Lima Lobo, A. V., Coutinho, F. R., Pfrimer, F. W. D., Lindino, C. A., Oyamada, M. S., & Martins, L. D. (2023). Low-Cost Water Quality Sensors for IoT: A Systematic Review. *Sensors*, 23(9), 4424. https://doi.org/10.3390/s23094424

[5] Teitelbaum, C. S., Ferraz, A., De La Cruz, S. E. W., Gilmour, M. E., & Brosnan, I. G. (2024). The potential of remote sensing for improved infectious disease ecology research and practice. *Proceedings of the Royal Society B: Biological Sciences*, 291(2037), 20241712. https://doi.org/10.1098/rspb.2024.1712

[6] WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP). (2025). *Progress on Household Drinking Water and Sanitation 2000–2024: Special Focus on Inequalities*. Geneva: World Health Organization and UNICEF.

[7] Xue, J-B., Xia, S., Wang, X-Y., Huang, L-L., Huang, L-Y., Hao, Y-W., Zhang, L-J., & Li, S-Z. (2023). Recognizing and monitoring infectious sources of schistosomiasis by developing deep learning models with high-resolution remote sensing images. *Infectious Diseases of Poverty*, 12, 10. https://doi.org/10.1186/s40249-023-01060-9

[8] Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. (2024). A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. *Journal of Marine Science and Engineering*, 12(1), 159. https://doi.org/10.3390/jmse12010159