A Foundational Dataset for the Predictive Prevention of Waterborne Disease

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose the creation of a new, open, and continuous dataset generated by a global network of autonomous, solar-powered smart buoys. Today, the fight against waterborne diseases like cholera and typhoid is fundamentally reactive, crippled by a critical data bottleneck: the lack of timely, high-resolution information on water quality. Manual sampling is slow, sparse, and expensive, meaning authorities only learn of a contamination event after people are already sick. By providing a continuous stream of molecular-level pathogen data fused with environmental metrics, this dataset will, for the first time, enable the AI community to build sophisticated, real-time forecasting models for disease outbreaks. Our vision is to catalyze a global, AI-powered early warning system that transforms public health from a reactive to a proactive science, preventing outbreaks and saving lives in the world's most vulnerable communities.

1 AI Task Definition: Multi-Scale Spatiotemporal Contamination Forecasting

The primary AI task is **predictive time-series forecasting and dynamic risk modeling**. The central scientific question is: *Can we accurately forecast the movement of contaminants and predict the likelihood of a waterborne disease outbreak in a specific location 24-72 hours in advance, based on real-time sensor data and environmental factors?*

This overarching goal is composed of three interconnected AI tasks:

- **Task 1: Time-Series Forecasting of Contaminant Levels.** Given the historical data from the buoy network, the task is to predict the future concentration of specific pathogens and pollutants at each sensor location. This is a classic sequence modeling problem well-suited for state-of-the-art models like **LSTMs, GRUs, and Transformer networks**, which can capture complex temporal dependencies and seasonalities.
- **Task 2: Geospatial Risk Mapping and Classification.** This task involves fusing the time-series predictions with static and dynamic environmental data (e.g., population density, rainfall anomalies, land use) to classify geographic sub-regions by their immediate outbreak risk (e.g., low, medium, high). Models like **XGBoost and Random Forest** will be trained to generate dynamic, hourly-updated risk maps that are intuitive for public health officials.
- **Task 3: Causal Inference and Driver Analysis.** Beyond prediction, it is critical to understand *why* an outbreak is likely. The task is to identify the key environmental and sensor-based factors driving a high-risk prediction. This will be accomplished using **Explainable AI (XAI) techniques like SHAP (SHapley Additive exPlanations)** to interpret model outputs, providing actionable insights for targeted interventions.

Submitted to the AI for Science workshop (NeurIPS 2025).

2 Proposed Dataset Schema:

Data Stream	Component	Details	Collection Freq.
1. Real-Time Buoy Data	Pathogen Biosensors	Graphene-based biosensors providing molecular-level detection of key pathogens like *E. coli, V. cholerae, S. typhi,* and chemical fingerprints for pollutants like PFAS.	15-minute intervals
	Water Quality Metrics	Standard multi-modal probes measuring pH, turbidity, conductivity, temperature, and oxidation-reduction potential (ORP).	15-minute intervals
	Geospatial Data	Precise GPS coordinates (latitude, longitude) for each buoy.	15-minute intervals
2. Integrated Contextual Data	Climate & Weather	Precipitation, temperature anomalies, humidity from NASA Giovanni & NOAA archives.	Synced Daily
	Socio-environmental	Population density (NASA SEDAC), land use maps, and infrastructure locations (bridges, industrial sites) to identify pollution sources.	Synced Weekly/Monthly
3. Ground-Truth Labels	Health Surveillance	Anonymized, location-tagged data on confirmed cases of waterborne diseases from partner public health organizations (e.g., PAHO, local Ministries of Health).	Synced as available

Scale and Scope: The project will launch with a 200-buoy pilot network in a high-risk watershed (e.g., a major river system in a cholera-endemic region). The long-term vision is to scale to thousands of units, creating a global, open-source repository for freshwater health.

Appendix

Alternate Dataset Variation: Fish Scales as Bio-Indicative Tissue Dataset

An alternative approach involves using fish scales as a bio-indicative tissue reflecting environmental conditions in subsequent growing seasons. In a recent study, fish scales from species such as common carp (*Cyprinus carpio L.*), chub (*Squalius cephalus*), and nase (*Chondrostoma nasus*) were found to be highly sensitive to the accumulation of metals like Mn, Ni, and Pb. In some parts of the scales, concentrations of these metals were up to ten times higher than in the fish's soft tissues, and these values significantly correlated with metal levels in liver and kidney tissue.

This suggests strong potential for developing *colorimetric methods with smartphone analysis* to detect and quantify such metals via scale-based biosensors.

Analytical Techniques

- **X-ray Fluorescence (XRF)** and **Inductively Coupled Plasma Optical Emission Spectrometry (ICP-OES)** have been employed to quantify elements such as Ca, K, Mg, Na, P, S, Al, Ba, Cu, Cr, Fe, Mn, Sr, and Zn in the scales of species like chub (*Squalius cephalus*) and nase (*Chondrostoma nasus*), demonstrating significant variability and species-specific accumulation patterns.
- **Scanning Electron Microscopy (SEM)** is used to characterize scale morphology, identifying structural differences that may influence element retention and accumulation.
- **Micro-XRF Mapping** provides qualitative elemental maps to pre-screen samples before quantitative ICP-OES analysis, enabling cost-effective sample selection.
- **Laser-Induced Breakdown Spectroscopy (LIBS)** has shown promise in detecting Fe and Pb in fish scales (e.g., *Salminus brasiliensis*, *Prochilodus lineatus*), outperforming atomic absorption spectroscopy in sensitivity—particularly for lead.

Environmental and Structural Insights

- Fish scales, as calcareous structures with growth rings, incorporate pollutants over time, enabling reconstruction of exposure history. Non-lethal sampling and repeated monitoring are possible.
- Metal accumulation in scales can occur within days of exposure and often correlates linearly with ambient water concentrations.
- Fish fins, like scales, can also bioaccumulate trace metals (e.g., As, Cu, Hg) and in some cases show correlations with muscle tissue levels, depending on species and metal.

Acceleration Potential: From Reactive Response to Proactive Prevention

This dataset will not be an incremental improvement; it will be a transformational catalyst for public health, environmental science, and AI research.

- **For Public Health: Revolutionizing Emergency Response.** Instead of reacting to hospital reports, health officials can use AI-powered dashboards to receive automated alerts about high-risk areas. This enables proactive, targeted interventions: distributing water purification tablets, launching public awareness campaigns, and pre-positioning medical teams *before* a single person gets sick.
- **For Environmental Science: A "Digital Twin" for Waterways.** The dataset network will create the first high-resolution, dynamic map of a river's health. Scientists can use this data to precisely identify and track pollution sources in real-time and provide policymakers with the quantitative evidence needed to enforce environmental regulations.
- **For AI Research: Catalyzing New Algorithmic Frontiers.** dataset will become a benchmark dataset for a new class of socio-environmental AI models. It will drive innovation in:

  - **Graph Neural Networks (GNNs):** Modeling entire river systems as graphs to predict how contaminants flow and disperse through the network.
  - **Physics-Informed Neural Networks (PINNs):** Integrating the physical laws of fluid dynamics into deep learning models to improve prediction accuracy.
  - **Epidemiology-Aware AI Models (EAAMs):** Developing agent-based models that simulate human behavior and its interaction with the dynamic contamination map to forecast disease spread.

Data-Creation Pathway: A Distributed, Autonomous Sensor Network

The data will be generated by a newly designed, open-standard hardware and software platform, ensuring transparency and community involvement.

- **Hardware - The "Smart Buoy":** Each unit is a self-sustaining, low-cost data collection platform.
  - *Sensing:* A modular sensor array featuring novel graphene-based biosensors for pathogens and standard probes for chemical/physical properties.
  - *Power:* A hybrid system using a primary solar panel and secondary energy harvesting (piezoelectric/triboelectric) from water motion for continuous, long-term, autonomous operation.
  - *Communication:* A low-power, long-range wireless module (e.g., LoRaWAN or cellular IoT) to transmit data to a central gateway.
- **Platform Architecture & Data Flow:**
  1. *Ingestion:* Data from the buoy network is transmitted to regional gateways.
  2. *Integration & Fusion:* A centralized, cloud-based platform ingests raw sensor data and automatically fuses it with contextual data streams from NASA, NOAA, and other archives.
  3. *Storage & Access:* The integrated dataset is stored in a time-series database and made publicly available via a simple, open **API**, ensuring it is FAIR (Findable, Accessible, Interoperable, and Reusable).
- **Deployment Strategy:** We will pursue a phased rollout in partnership with a local river authority and a public health NGO to ensure the pilot program is grounded in real-world needs.

Hardware Specifications

Our strategy is designed for cost-effective generation at scale, with a clear path from a pilot project to a global utility.

Buoy-Based E. coli Detection System Overview

The current version of the buoy tested utilizes an optochemical reaction that changes color in the presence of *E. coli*, with a camera mounted on top to capture and relay the visual information. The system involves two main vials used for the biochemical detection of coliforms.

First Vial: Lauryl Tryptose Broth (LTB) Powder

This vial contains a dehydrated powder formulation of Lauryl Tryptose Broth, designed to be reconstituted with 10 mL of water obtained from a pre-concentrated 100 mL water sample. The standard composition per liter of LTB is:

Final pH: 6.8 ± 0.2 at 25°C.

Function: This medium is commonly used for the detection of coliform bacteria in water and food samples.

Component	Concentration (g/L)
Tryptose	20.0
Lactose	5.0
Sodium Chloride (NaCl)	5.0
Dipotassium Hydrogen Phosphate (K_2HPO_4)	2.75
Potassium Dihydrogen Phosphate (KH_2PO_4)	2.75
Sodium Lauryl Sulfate (SLS)	0.1

Second Vial: Gel-Based Detection Matrix

This vial contains a gel matrix (100–1000 µL) composed of the following:

- **Agar or Purified Agar:** 2–20 mg
  Serves as a gelling agent to provide structural integrity.
- **Triton X-100 in Tris-HCl Buffer (10–50 mM, pH 7.5–8.5):** 25–100 µL
  Triton X-100 is a non-ionic surfactant that lyses cells, releasing intracellular enzymes. Tris-HCl buffer maintains pH stability.
- **Lauryl Tryptose Broth Solution:** 100–400 µL
  Provides nutrients to support coliform bacterial growth during incubation.
- **Chemical Solution/Mixture:** 50–200 µL, containing:
  - **6-Chloro-3-indolyl-β-D-galactopyranoside (Red-Gal):** 20–90 mg
    A chromogenic substrate for β-galactosidase. Upon enzymatic cleavage, a colored product is formed, enabling visual detection.
  - **N,N-Dimethylformamide (DMF):** 1000–2000 µL
    Solvent for Red-Gal, which is sparingly soluble in water.
  - **Deionized Water:** 1000–2000 µL
    Used to dilute the chemical mixture to the desired concentration.

Function: This matrix visually indicates coliform presence by detecting β-galactosidase activity, which cleaves Red-Gal to produce a colored compound.

Data Processing and Transmission

An edge-computing microcontroller processes sensor data and transmits it via a LoRaWAN/IoT network. Collected data is streamed to an online system for real-time processing and visualization as an AI-assisted geospatial contamination map.

Additional Features:

- Parameters such as turbidity, pH, conductivity, and oxidation-reduction potential (ORP) are simultaneously recorded.
- The AI model analyzes this dataset to predict contamination hotspots and pollutant dispersion patterns along the river's path through villages, urban centers, and industrial zones.

Proposed Biosensor Methodology

The next step is to create biosensors via the immobilization of enzymes that target analytes released during the respiration of *Salmonella*, *Shigella*, *Cholera*, and *E. coli*. When these bacteria respire, they produce analytes that, upon contacting specific enzymes, generate a varying electric current measurable with an ammeter to determine bacterial presence.

Graphene on a plastic film is submerged in a $5\,\mathrm{mM}$ solution of 1-pyrenebutyric acid (PBA) dissolved in DMSO and left at room temperature. After two hours, the substrates are removed and rinsed with distilled water, bonding the PBA onto the graphene through π-π interactions.

Enzyme immobilization: Glucose Oxidase, Glycerol-3-Phosphate Oxidase, Galactose Oxidase, and Lactate Oxidase are immobilized onto the PBA-functionalized graphene. First, $24\,\mu\mathrm{L}$ of distilled

water is added to the anhydrous enzyme, and the solution is ultrasonicated for $180\,\mathrm{s}$. This process is repeated for each enzyme. One aliquot of each enzyme is micro-pipetted onto a separate graphene substrate and stored at $7\,^{\circ}\mathrm{C}$ overnight to complete immobilization.

Conductive paint terminals are applied on opposite, parallel sides of the graphene biosensor to facilitate electrical measurements.

Sensor for PFAS Detection

The sensor is built on a **lateral-flow paper-based platform**—a format similar to that used in COVID-19 and pregnancy tests. Unlike antibody-based visual readouts, this sensor uses a **conductive polymer** to detect PFAS (per- and polyfluoroalkyl substances) through changes in electrical resistance.

Core Technical Components

Polymer: Polyaniline

The active sensing element is **polyaniline**, a conductive polymer that can reversibly switch between semiconducting and conducting states when protons interact with it.

Substrate: Nitrocellulose Paper

Polyaniline is deposited onto a **nitrocellulose strip**, which acts as a structural substrate. This material allows for **fluid wicking** and ensures the platform remains lightweight and portable.

Surfactant Coating

A **surfactant layer** coats the strip and facilitates the extraction of **acidic PFAS molecules**, including:

- PFOA (Perfluorooctanoic acid)
- PFBA (Perfluorobutanoic acid)

These molecules are drawn from a small droplet of water into the sensing region.

Detection Mechanism

PFAS-Driven Proton Transfer

When acidic PFAS enter the sensor strip, **protons are transferred into the polyaniline**, switching it from a semiconducting to a conducting state. This transition causes a **drop in electrical resistance**, which forms the basis for detection.

Quantitative Readout

Electrodes embedded in the device measure the change in resistance, yielding a quantitative signal that corresponds to the PFAS concentration. This electrical signal can be:

- **Read via an ammeter**
- **Transmitted to a smartphone** or external device for data processing and display

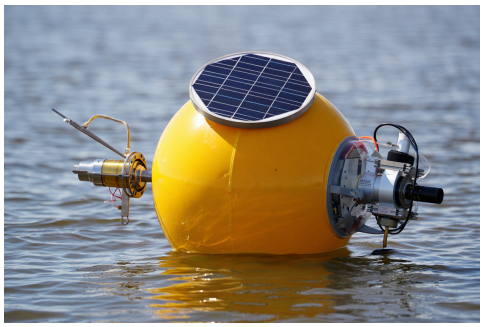

Figure 1: Proposed Buoy System

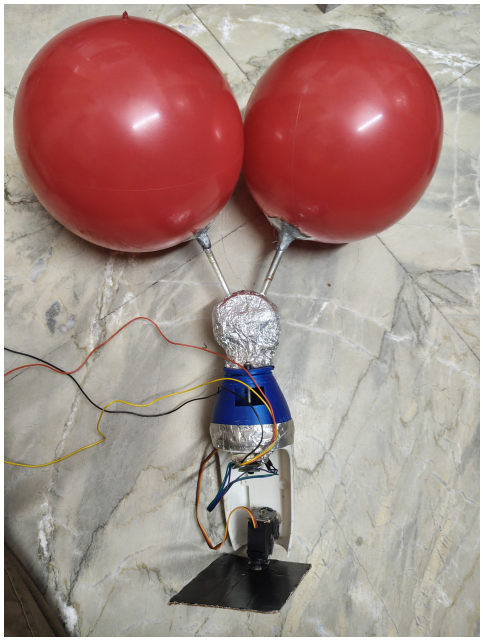

Figure 2: Early prototype version carrying the red-gal chemical test for e.coli

References

1 United Nations Environment Programme. (2021, March 22). Globally, 3 billion people at health risk due to scarce data on water quality. UNEP News. https://www.unep.org/news-and-stories/story/globally-3-billion-people-health-risk-due-scarce-data-water-quality

2 World Health Organization. (2022, March 22). Drinking-water. WHO Factsheet on Global Water Contamination. https://www.who.int/news-room/fact-sheets/detail/drinking-water

3 World Bank. (2025, February 28). World Water Day: Two billion people still lack access to safely managed water. https://blogs.worldbank.org/en/opendata/world-water-day-two-billion-people-still-lack-access-safely-managed-water

4 Shah, A., Arjunan, A., Baroutaji, A., & Zakharova, J. (2023). A review of physicochemical and biological contaminants in drinking water and their impacts on human health. *Water Science and Engineering, 16*(4), 333–344. https://doi.org/10.1016/j.wse.2023.04.003

5 SIWI. (2017, August 30). Ryan Thorpe and Rachel Chang from the USA win 2017 Stockholm Junior Water Prize. *Stockholm International Water Institute*. https://siwi.org/latest/ryan-thorpe-rachel-chang-usa-win-2017-stockholm-junior-water-prize/

6 Odobašić, A., Šestan, I., & Begić, S. (2019). Biosensors for determination of heavy metals in waters. *IntechOpen*. https://doi.org/10.5772/intechopen.84139

7 Hassan, M., Zhao, Y., & Zughaier, S. M. (2024). Recent advances in bacterial detection using surface-enhanced Raman scattering. *Biosensors, 14*(8), 375. https://doi.org/10.3390/bios14080375

8 Payne, T. D., Klawa, S. J., Jian, T., Wang, Q., Kim, S. H., Freeman, R., & Schultz, Z. D. (2023). From the lab to the field: handheld surface enhanced Raman spectroscopy (SERS) detection of viral proteins. *Sensors & diagnostics, 2*(6), 1483–1491. https://doi.org/10.1039/d3sd00111c

9 Chen, S.-E., Yang, R.-Y., Qiu, Z.-H., & Wu, C.-C. (2021). A Piezoelectric Wave Energy Harvester Using Plucking-Driven and Frequency Up-Conversion Mechanism. *Energies, 14*(24), 8441. https://doi.org/10.3390/en14248441

10 Xie, X. D., Wang, Q., & Wu, N. (2014). Energy harvesting from transverse ocean waves by a piezoelectric plate. *International Journal of Engineering Science, 81*, 41–48. https://doi.org/10.1016/j.ijengsci.2014.04.003

11 Nasr, N., Shafi, M., Zhao, T., Ali, R., Ahmad, I., Khan, M., Deifalla, A., Ragab, A. E., & Ansari, M. Z. (2024). A two-fold SPR-SERS sensor utilizing gold nanoparticles and graphene thin membrane as a spacer in a 3D composite structure. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 304*, 123331. https://doi.org/10.1016/j.saa.2023.123331

12 Wei, Q., Nagi, R., Sadeghi, K., Feng, S., Yan, E., Ki, S. J., Caire, R., Tseng, D., & Ozcan, A. (2014). Detection and spatial mapping of mercury contamination in water samples using a smart-phone. *ACS Nano, 8*(2), 1121–1129. https://doi.org/10.1021/nn406571t

13 Marie Adier, Anne-Marie Jurdyc, Charlotte Hurel, François Goutaland, Jean-Yves Michalon, et al. (2023). Toward surface-enhanced Raman scattering using electroless substrate for trace arsenic detection and speciation. *Journal of Applied Physics, 133*(7), 073103. https://doi.org/10.1063/5.0126372

14 Park, S., Gordon, C.T., & Swager, T.M. (2024). Resistivity detection of perfluoroalkyl substances with fluorous polyaniline in an electrical lateral flow sensor. *Proceedings of the National Academy of Sciences, 121*(12), e2317300121. https://doi.org/10.1073/pnas.2317300121

15 Viana, L. F., Súarez, Y. R., Cardoso, C. A. L., Lima, S. M., da Cunha Andrade, L. H., & Lima-Junior, S. E. (2019). Use of fish scales in environmental monitoring by the application of Laser-Induced Breakdown Spectroscopy (LIBS). *Chemosphere, 228*, 258–263. https://doi.org/10.1016/j.chemosphere.2019.04.070

16 Marić, B., Rašković, B., Babović, N., Hegediš, A., Spasić, M., Lenhardt, M., & Poleksić, V. (2023). Prospects of fish scale and fin samples usage for nonlethal monitoring of trace element contamination in freshwater fish. *Knowledge and Management of Aquatic Ecosystems, 424*, 12. https://doi.org/10.1051/kmae/2023010

17 Varga, M., Szakál, A., Kálmán, E., Harangi, S., Baranyai, E., & Nagy, S. A. (2025). Fish Scales as a Non-Invasive Method for Monitoring Trace and Macroelement Pollution in Rivers. *Animals, 15*(6), 943. https://doi.org/10.3390/ani15060943

18 Vaid, V., & Hundal, S. S. (2024). Ultrastructural Studies to Evaluate Fish Scales as Indicators of Heavy Metals in *Labeo rohita*. *Journal of Soil Salinity and Water Quality, 16*(1), 89–96. https://doi.org/10.56093/JSSWQ.v16i1.148991

19 Jain, S., Chattopadhyay, S., Jackeray, R., Abid, C. K. V. Z., Kohli, G. S., & Singh, H. (2017). DipTest: A litmus test for E. coli detection in water. *PLoS ONE, 12*(9), e0183234. https://doi.org/10.1371/journal.pone.0183234

20 Mallmann, W. L., & Darby, C. W. (1941). A new medium for the enumeration of coliform organisms. *American Journal of Public Health, 31*(2), 127–132. https://doi.org/10.2105/AJPH.31.2.127

21 Li, Y., Wang, Z., Sun, L., Liu, L., Xu, C., & Kuang, H. (2019). Colorimetric Detection of *Escherichia coli* Based on the Enzyme-Induced Metallization of Gold Nanorods. *ACS Applied Materials Interfaces, 11*(46), 43382–43389. https://doi.org/10.1021/acsami.9b16758

22 Earthface Private Limited. (2025). Padma Bio: E. coli Water Testing Kit [Product information]. https://www.earthface.in/product/padma-bio