Large-Language-Model-Assisted Microbial Source Tracking Reveals Mechanisms of Tropical Cyclone Impacts on Microbial Water Quality in Coastal Environments

Chamteut Oh

chamteutoh@ufl.edu

University of Florida https://orcid.org/0000-0002-1239-6482

Van-Khuong Trinh

University of Florida

Palash Sashittal

Virginia Tech

Shin-Ah Lee

University of Florida

Amanda Chappel

University of Florida

Megan Sanford

University of Florida

David Tomasko

Sarasota Bay Estuary Program

Edward Sherwood

Tampa Bay Estuary Program https://orcid.org/0000-0001-5330-302X

Jennifer Hecker

Coastal & Heartland National Estuary Partnership

Abbey Tyrna

Suncoast Water Keeper

Sungyoon Jung

University of Florida

Elise Morrison

University of Florida https://orcid.org/0000-0003-2569-4259

Article

Keywords: Tropical cyclone, microbial source tracking, large language models, coastal environments, pathogens, fecal indicator bacteria, antimicrobial resistance genes

Posted Date: February 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8765875/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Extreme weather events increasingly threaten coastal water quality, yet the mechanisms by which tropical cyclones impair microbial water quality remain poorly quantified. We developed a Large-Language-Model-Assisted Microbial Source Tracking (LAMST) framework to trace the origins of microbial threats and applied it to coastal waters impacted by Hurricane Milton along Florida's Gulf Coast. LAMST integrates 16S rRNA sequencing with species-level metadata from the NCBI BioSample database to classify taxa as marine, terrestrial, or wastewater in origin. Across 30 sites and three time points (1 week to 7 months post-storm), twelve microbial analytes, including fecal indicator bacteria (FIB), pathogens, and antimicrobial resistance genes (ARGs), were quantified. Colored dissolved organic matter as well as carbon and nitrogen stable isotopes in particulate material were characterized for the same samples and independently corroborated LAMST's terrestrial and marine microbial classifications. Increases in terrestrial and marine bacterial counts in coastal waters indicated concurrent mobilization of land-derived inputs and marine sources following the hurricane. Site fixed-effects regressions showed terrestrial and wastewater bacteria were strongly associated with enterococci and two ARGs (i.e., *sul2* and *tetA*), whereas marine bacteria correlated with *Vibrio parahaemolyticus*. LAMST provides a quantitative, generalizable framework for source-resolved microbial tracking without reliance on regional reference sequences.

Synopsis

Large Language Models enable microbial source tracking without regional reference data. This approach reveals how Hurricane Milton impacted microbial water quality along Florida's Gulf Coast.

1. Introduction

Tropical cyclones involve strong winds and often heavy rainfall. These forces degrade microbial water quality in coastal environments through two primary pathways. First, wind stress and wave energy drive onshore water movement and storm surge, disturbing seabed sediments and resuspending particles.[1,2] Marine and estuarine surface sediments act as long-lived reservoirs of fecal indicator bacteria (FIB), pathogens, and antibiotic resistance genes (ARGs).[3–11] Laboratory experiments have shown that even low-energy conditions can mobilize bacteria from sediments into the overlying water.[12] Second, heavy rainfall and the return flow of storm surge water cause flooding and mobilize contaminants, including stormwater and wastewater overflows, that transport microbial pollutants to coastal environments.[13–18] Given these processes, tropical cyclones substantially impact coastal ecosystems and microbial water quality—a critical concern as nearly 40% of the global population lives within 100 km of a coastline.[19]

Although cyclone-induced impairments of coastal water quality are well documented, the quantitative contributions of the two key mechanisms—marine sediment resuspension and inland discharge—remain unclear. Part of this uncertainty stems from large event-to-event differences in storm intensity and rainfall totals, which are fundamental drivers of contaminant transport. Coastal water quality is also

highly dynamic in space and time.[20] Hydrodynamics and river inputs vary rapidly, and the transport of microbial contaminants such as fecal indicator bacteria (FIB), pathogens, and antimicrobial resistance genes (ARGs) is difficult to resolve. Moreover, microbial taxa differ in salinity tolerance and persistence, adding another layer of complexity to their fate and transport. As tropical cyclone events become more frequent, clarifying how they impair coastal water quality is essential for protecting public and environmental health.

While either mechanism can dominate depending on environmental conditions, microbial source tracking (MST) is a valuable tool for identifying sources of microbial contaminants.[21] MST generally follows two complementary approaches.[22] The first employs host-specific proxy markers—genetic targets that occur predominantly in a given host group (e.g., human HF183, ruminant Rum2Bac, avian markers)—which are quantified by qPCR or dPCR.[23−25] This approach is rapid, relatively inexpensive, and directly quantitative.[26,27] However, each marker detects only its target host, and marker persistence may differ from that of pathogens or ARGs, leading to potential mismatches with actual contaminant distributions or the omission of atypical sources.[28,29] The second approach involves profiling entire microbial communities (e.g., 16S rRNA sequencing or metagenomes) and comparing them to reference sequences obtained in upstream water environments using source-apportionment tools such as SourceTracker or FEAST.[22,30−32] This method can apportion multiple sources simultaneously and provide ecological context, but it depends on robust and representative regional reference sequences.[33,34] Because microbial communities vary spatiotemporally, the reference sequences must be frequently updated to remain valid. However, in practice, many regions or contamination scenarios (particularly during unexpected weather events such as tropical cyclones) lack suitable references, limiting the applicability of this approach.[35]

Large language models (LLMs) have rapidly improved in both accuracy and utility for research. Beyond simply supporting writing and coding, they have become essential tools for data mining tasks such as generalizing from limited examples and extracting entities and relationships from unstructured data. In environmental science and engineering, two applications stand out. First, LLMs excel at condensing large bodies of literature into clear syntheses while highlighting differences across studies. This enables researchers to detect broad temporal or spatial trends, identify robust findings, and pinpoint knowledge gaps or priorities for further study. Recent review papers have applied this approach to streamline evidence mapping.[36−38] Second, LLMs are effective at transforming unstructured prose into structured datasets. Key details in scientific papers or permits—such as analytes, detection methods, or site information—can be extracted into tables for direct comparison. This has facilitated meta-analyses and data harmonization, for example, by compiling ARG concentrations, methods, and locations from publications or pollutant limits from permits.[39−41] The outcome is faster, more consistent data assembly to support modeling and decision-making.

Here, we developed an LLM-Assisted Microbial Source Tracking (LAMST) approach to trace sources of microbial contamination—FIB, pathogens, and ARGs—following tropical-cyclone–driven storm surge and

heavy rainfall. We collected time-series coastal water samples from 30 sites spanning about 150 km of Florida's Gulf Coast after Hurricanes Helene and Milton (77 observations). Each sample was quantified for 11 targets—total bacteria; pathogens (*E. coli*, *Pseudomonas*, *Vibrio parahaemolyticus*, *Vibrio vulnificus*); and ARGs (sul1, sul2, qnrS, tetA, floR, aadA)—by digital PCR (dPCR), as well as enterococci by a culture-dependent assay (EPA Method 1600). We also profiled the bacterial community using 16S rRNA gene amplicon sequencing. LAMST links observed taxa to curated National Center for Biotechnology Information (NCBI) metadata and assigns each to its most likely origin (marine, terrestrial, or wastewater), yielding quantitative, organism-resolved source contributions. Source contributions were independently corroborated with geochemical measures of colored dissolved organic matter (CDOM) as well as carbon and nitrogen isotopes in particulate material. We systematically validated LAMST and show how Hurricane Milton altered microbial water quality along Florida's Gulf Coast by disentangling the relative roles of wastewater overflow, terrestrial runoff, and marine sediment resuspension.

2. Results

2.1. Impact of Hurricane Milton on coastal microbial water quality

We analyzed water samples from 30 hurricane-impacted sites at three time points (1 week, 2 weeks, and 7 months after Hurricane Milton). Overall, we found the high concentrations of analytes in sampling sites along the Manatee River (TB1, TB2, MR1, MR2, MR3, MR4, and MR5), which were closest to the trajectory of Hurricane Milton (Fig. 3). This trend aligns with the exceptionally high number of reported sanitary sewer overflows in September 2024 (n = 22; Hurricane Helene landfall) and October 2024 (n = 16; Hurricane Milton landfall) within the Manatee River watershed, compared to the monthly median of one overflow event during the past seven years (**Fig. S4**).

Enterococci concentrations ranged from 0 to 390 CFU/100mL one week after the hurricane. Average enterococci concentrations declined rapidly, from 43 CFU/mL at 1 week, to 11 CFU/mL at 2 weeks, and 0.3 CFU/mL at 7 months, indicating rapid recovery of microbial water quality within the first few weeks. According to Florida regulations for beach water quality (0–35 CFU/mL = good, 36–70 = moderate, > 71 = poor),[42] five of the 30 sites showed poor water quality, three showed moderate water quality, and 22 showed good water quality one week after the hurricane. Two weeks after the hurricane, only one out of 17 sites showed moderate water quality, while the other 16 sites showed good quality. Overall, the extent of water quality impairment was less severe than that observed after other hurricanes.[43] If relying solely on enterococci results, the hurricane-impacted coastal water would have been considered safe within a few weeks.

However, other water quality parameters did not fully agree with the enterococci results. Using dPCR, we analyzed 11 analytes: total bacteria (16S rRNA), six ARGs (aadA, sul1, sul2, floR, qnrS, tetA), and four pathogens (*E. coli, Pseudomonas, V. vulnificus*, and *V. parahaemolyticus*). Total bacteria, sul1, sul2, qnrS,

and *Pseudomonas* were consistently detected across all observations, while tetA, floR, aadA, and *E. coli* were below the detection limit in some samples, mostly at the 7-month sampling interval. Additionally, we analyzed six of the most impacted samples for four extra ARGs (metA, tetW, tetO, and blaTEM) and five additional pathogens (*Klebsiella rcSA* gene, *Staphylococcus cap* gene, *Aeromonas ahal* gene, *Legionella mip* gene, and *Clostridium cpa* gene). All of these were below detection limits, so further analysis was not performed on the remaining samples.

We conducted paired t-tests or Wilcoxon signed rank test on samples from the same sites across different time points depending on the data normality (**Table S4**). Comparisons between 1-week and 7-month samples showed that concentrations of all 12 analytes were higher 1-week post-storm, indicating that Hurricane Milton impaired microbial water quality (Fig. 1E-P). However, comparisons between 1-week and 2-week samples showed that eight analytes (all except sul1, sul2, and *V. vulnificus*) were not significantly different between these two time points. Based on these water quality measurements, Hurricane Milton impaired overall microbial water quality, but recovery trajectories differed depending on the indicator used.

2.2. Application of the Large-Language-Model-Assisted Microbial Source Tracking (LAMST)

We developed the Large-Language-Model-Assisted Microbial Source Tracking (LAMST) framework to investigate how hurricanes impair microbial water quality at our study sites. This approach analyzes microbial species in each sample using 16S rRNA sequencing and infers their most likely environment of origin leveraging metadata from the NCBI BioSample database (Fig. 2A). To validate the framework, we first applied it to reference sequences deposited in the NCBI Sequence Read Archive (SRA) (Fig. 2B). The reference sequences were derived from samples collected from three distinct environments: marine, terrestrial, and wastewater. Within each group, the relative abundance of predicted bacterial origin varied, with standard deviations ranging from 0.03 (wastewater bacteria in terrestrial samples) to 0.22 (marine bacteria in marine samples). These variations likely originated from different sampling sites collected by different research groups for different research purposes. Nevertheless, the sampling sites (marine, terrestrial, or wastewater environments) were significantly associated with the most likely microbial origin inferred by LAMST. For example, marine samples had a mean relative abundance of 0.49 for marine-associated bacteria, significantly higher than that of terrestrial ($p < 0.001$) and wastewater samples ($p < 0.001$). Terrestrial samples had a mean relative abundance of 0.80 for terrestrial bacteria, significantly higher than that of marine ($p < 0.001$) and wastewater samples ($p < 0.05$). Similarly, wastewater samples had a mean relative abundance of 0.21 for wastewater bacteria, significantly higher than that of marine ($p < 0.001$) and terrestrial samples ($p < 0.05$).

Geochemical measurements (CDOM, stable isotopes) were used to independently corroborate classifications from LAMST. One and two weeks after the hurricane, the relative abundance of terrestrial bacteria were significantly and positively correlated with CDOM (a254; one week, $R = 0.81$, $p < 0.05$; two weeks $R = 0.92$, $p < 0.05$) and significantly and inversely correlated with salinity (one week, R=-0.68, $p <$

0.05; two weeks R=-0.77, $p < 0.05$). Terrestrial bacteria were also significantly and inversely correlated with carbon stable isotope values one week after the hurricane (R=-0.52, $p < 0.05$; R=-0.43, $p < 0.05$). All three parameters independently corroborate LAMST's classification of terrestrial bacteria (Fig. 3A,3B, and 3C). The relative abundance of wastewater bacteria also showed the same significant relationships with CDOM, salinity, and carbon stable isotopes one and two weeks after the hurricane (Fig. 3D, 3E, and 3F), indicating that wastewater microbes were predominantly derived from freshwater sources after the hurricane. High nitrogen stable isotope values can be associated with wastewater[44–46]. Interestingly, wastewater bacteria were only significantly positively correlated with nitrogen stable isotopes two weeks after the hurricane (R = 0.68, $p < 0.05$) (**Fig. S5**). The lack of a relationship with wastewater bacteria at the one week and control timepoints are likely due to different nitrogen sources and processing at those times. For the same geochemical parameters, the relative abundance of marine bacteria showed significant, but opposite relationships as terrestrial/wastewater bacteria, also independently corroborating LAMST's classification of marine bacteria (Fig. 3G, 3H, and 3I). However, during the control timepoint, terrestrial, marine, and wastewater bacteria did not exhibit any significant relationship with the geochemical proxies measured, likely due to differences in estuarine processes under non-hurricane conditions.

We then applied LAMST to the water samples collected from the middle and lower estuaries in this study. The middle estuary sites are strongly influenced by land and riverine inputs, whereas the lower estuary sites are directly connected to the open ocean, where inland discharges are more readily diluted (**Table S1**). One week after Hurricane Milton, the middle estuary exhibited significantly higher relative abundances of terrestrial (mean values 0.61 vs 0.23) and wastewater bacteria (0.06 vs 0.04) and significantly lower relative abundance of marine bacteria (0.33 vs 0.73) compared with the lower estuary (Fig. 2C), reflecting the stronger impact of inland discharge at the middle estuary. In addition, the middle estuary one-week post-hurricane contained significantly more terrestrial (0.61 vs 0.16) and wastewater bacteria (0.06 vs 0.04) than during the control season. As a result, the relative abundance of marine bacteria was significantly reduced from 0.80 (at the non-storm sampling event) to 0.33 (1 week after Hurricane Milton). Likewise, the lower estuary showed elevated terrestrial bacteria (0.23 vs 0.14) compared to the non-storm sampling event. These findings indicate that the hurricane transported inland bacteria to coastal water environments, even to the lower estuary. By contrast, during the control season (Fig. 2D), no significant differences in terrestrial (0.16 vs 0.14) or wastewater bacteria (0.04 vs 0.05) were observed between middle and lower estuary sites, suggesting that under normal conditions, inland discharge has minimal influence on coastal microbial composition.

2.3. Revealing the mechanisms by which Hurricane Milton impaired microbial water quality in the coastal environment.

We examined mechanisms by which Hurricane Milton impaired microbial water quality in the coastal environment. Figure 4 reports fixed-effects within-site tests evaluating whether increases in absolute terrestrial, wastewater, and marine bacterial loads (the first three rows in Fig. 4) are associated with changes in total bacteria and with individual analytes, including FIB, pathogens, and ARGs. First, total

bacteria count rose across the study area because both terrestrial and marine bacteria increased (elasticities of 43% and 51%, respectively, in Fig. 4). However, each microbial threat (FIB, pathogens, ARGs) is associated with different components of bacterial origins. For example, higher wastewater-origin load was associated with significant increases in enterococci, *Sul2* and *tetA* gene with positive elasticities ($p < 0.05$). This pattern supports wastewater inputs as an important storm-driven source of these microbial threats to the coastal environment. Terrestrial-origin bacteria were also significant predictors for enterococci ($p < 0.05$) and *sul2* gene ($p < 0.001$). These findings are consistent with the idea that enterococci and ARGs are widespread in the natural environments and can be transported through non-sewered runoff. In Florida, the cattle industry uses sulfonamide antibiotics for the control and prevention of diseases for livestock.[47,48] Consequently, watershed delivery of sulfonamide resistance, such as Sul2, is plausible in runoff-impacted areas with cattle. As of 2024, more than 60,000 head of cattle were raised in Charlotte, Manatee, and Sarasota Counties, the three counties that encompass most of our study area.[49] Moreover, the marine-origin bacteria were significantly associated with *V. parahaemolyticus*, which supports the statistical analysis in Fig. 4, as this *vibrio* species is known to be a marine pathogen.[50–52]

3. Discussion

3.1. LAMST is a widely applicable approach for quantitative microbial contaminant source tracking

Tropical cyclones impair coastal microbial water quality through the resuspension of contaminated marine sediments and the discharge of stormwater and wastewater overflows that transport microbial contaminants from inland sources.[3–11,14–18] Although two MST approaches—host-specific proxy markers and community profiling against region-specific references—are well established, [22–25,30–32] they are impractical for large-scale, unexpected events such as hurricanes, where contaminants originate from diverse sources and suitable reference datasets are lacking.

This study develops the LAMST framework, a quantitative approach for tracing microbial contaminants. The LAMST framework overcomes the limitations of existing MST strategies while retaining their strengths. Unlike proxy-based methods, it analyzes the entire microbial community, providing a more accurate representation of the actual distribution of microbial contaminants. Moreover, unlike reference-based approaches, it does not depend on spatiotemporally matched datasets; instead, it infers microbial origins using publicly available metadata, enabling quantitative source apportionment without regional reference sequences. In the case of the current study, reference sequences should have been from terrestrial runoff, wastewater effluent, and marine sediments, which are practically impossible to obtain right before Hurricane Milton made landfall.[43] As a result, this approach is broadly applicable across environments and provides robust insights into the contributions of multiple microbial sources to coastal contamination.

We applied LAMST to coastal waters affected by Hurricane Milton and found that both terrestrial and marine bacteria increased in absolute abundance following the storm. These results align with reports that Hurricane Milton caused extensive inland runoff and untreated wastewater releases due to power outages and sewer overflows, likely enhancing land-derived microbial discharges.[53,54] In addition, strong winds and wave energy generated onshore water movement, corroborated by a sea-level setup that reached up to 3 m above ground level near Venice Beach, Florida, that supports enhanced resuspension of marine sediments.[55] We further quantified how each source—and their interactions—affected fecal indicator bacteria (FIB), pathogens, and antimicrobial resistance genes (ARGs) across the estuarine gradient. The results indicate that terrestrial and wastewater bacteria can serve as vectors through which ARGs establish within coastal microbiomes, increasing public health risks through recreation and seafood consumption,[56,57]

Practical measures include stormwater detention/retention (e.g., urban ponds, green infrastructure, regional stormwater facilities, etc.) to dampen runoff pulses; sewer rehabilitation, emergency backup power and overflow controls to limit untreated releases; enhancing wastewater-treatment resilience, including contingency disinfection during bypass events; and agricultural antibiotic stewardship to reduce watershed ARG loads.[58–60] On the receiving-water side, shoreline restoration that attenuates wave energy can curb sediment resuspension.[61]

3.2 Fecal indicator bacteria may not be the ideal proxy for microbial water quality in the coastal environments impacted by tropical cyclones

The LAMST approach revealed diverse sources of microbial contaminants in coastal waters impacted by tropical cyclones. We also found that FIB has two major limitations as proxies for assessing dynamic changes in microbial water quality in cyclone-impacted coastal environments. First, the enterococci signal (CFU/100 mL) declined rapidly from week 1 to week 2 post-storm (Fig. 1F). Because enterococci is the primary regulatory metric for coastal waters, this pattern could be mistaken for rapid water-quality recovery. In contrast, allochthonous taxa (e.g., *E. coli*, *Pseudomonas*), indigenous pathogens (e.g., *Vibrio spp.*), and multiple ARGs remained elevated and showed little change over the same period. In particular, ARGs may exert longer-term impacts because they can transfer to indigenous microbes via mobile genetic elements (MGEs—plasmids, integrons, transposons) and persist even after donor populations wane.[62,63] Such transfer could involve indigenous pathogens detected in our samples (e.g., *V. parahaemolyticus, V. vulnificus*), potentially increasing treatment challenges and clinical risk if they acquire clinically relevant ARGs.[18,64,65]

Second, it failed to capture broader microbial-contaminant signals. In our data, enterococci in CFU per 100 mL was significantly correlated only with two analytes (*sul1* and *sul2* genes), and *E. coli* in gc/L tracked with three analytes (*sul1*, *sul2*, and *qnrS* genes). One microbial genus (as in enterococci) or species (as in *E. coli*) cannot fully represent the broad range of coastal microbial contaminants, particularly ARGs, because ARGs are present across diverse bacterial species with different persistence in coastal waters.[66–68] Because of the wider presence of ARGs in various microbial species, ARGs may

provide a more informative picture of the overall microbial water quality in the coastal environments impacted by inland discharge. For example, increases in *sul1* and *sul2* coincided with rises in two FIB (*E. coli* and enterococci) and four ARGs (*qnrS*, *sul1* or *sul2*, and *floR* genes), and increases in *qnrS* likewise accompanied rises in multiple microbial threats (*E. coli*, *sul1*, *sul2*, qnrS, *tetA*, and *floR* genes).

3.3. Limitations and future work

This work has several limitations which highlight avenues for future work. Our Large-Language-Model-Assisted Microbial Source Tracking (LAMST) approach relies on species-level metadata in the NCBI BioSample database, which is sparse for many taxa. In this study, we queried 1,944 bacterial species; 292 species (~ 15%) had no metadata, and about half had two or fewer metadata records. However, low metadata frequency does not imply low environmental abundance. For example, *Fischerella indica* was the most dominant taxon at site ULB2 one week after Hurricane Milton (~ 31% relative abundance), yet it lacks BioSample metadata and was therefore labeled “insufficient” and excluded from marine/terrestrial/wastewater origin assignment. Depending on its true habitat, source-composition estimates for that sample could shift substantially. A second limitation is that LAMST assigns a single origin per species even though many taxa occur across multiple environments. BioSample XML often contains conflicting habitat terms. For instance, *Clostridium pasteurianum* metadata includes both “soil” and “wastewater treatment,” and LAMST assigned it to “terrestrial” because “soil” appeared more frequently (4 vs. 1). Likewise, *Bauldia litoralis* had two “marine” and one “freshwater lake” entries and was labeled “marine.” Because sampling is generally easier on land than at sea, database records may be biased toward terrestrial habitats. The current LAMST implementation does not correct for such bias or for multi-habitat taxa.

To address the two limitations, LAMST can be improved in two ways. First, we can broaden the evidence base beyond NCBI BioSample by ingesting structured and semi-structured sources (e.g., curated repositories, environmental monitoring reports, and peer-reviewed articles). Habitat cues can be extracted at the sentence level using a controlled vocabulary (e.g., isolated from, collected at, salinity/latitude constraints) with provenance tracking and confidence scores. Second, we can complement the LLM layer with an interpretable natural language processing approach. For example, word-embedding models can be employed to cluster habitat-related metadata and identify the most representative class (‘marine’, ‘terrestrial’ or ‘wastewater’) for each sample. This enables probabilistic multi-label classification that estimates the likelihood that each taxon originates from wastewater/animal or terrestrial/aquatic environments, based on weighted keyword evidence rather than forcing a single label. Such probabilistic classifiers can also down-weight duplicated records, apply temporal or quality-based weights, and propagate uncertainties into sample-level source proportions. Together, richer evidence and probabilistic, multi-origin attribution would mitigate metadata sparsity and habitat conflicts and make LAMST more robust for event-driven coastal studies.

4. Materials and methods

4.1. Site description and sample collection

We conducted a sampling campaign along Florida's Gulf Coast, spanning about 150 km and encompassing the Manatee River, Sarasota Bay, Lemon Bay, and the Peace River. These areas were impacted by two major hurricanes (Helene and Milton) in 2024. Hurricane Helene and Milton made landfall on September 26, 2024, and October 9, 2024, respectively (Fig. 1A). According to the Sarasota Bay Estuary Program, in the sampled region, Hurricane Helene reached top wind speeds exceeding 33.1 m/s, produced a storm surge of 2.1 m, and brought 64 mm of rainfall. Hurricane Milton caused local wind speeds up to 45.6 m/s, a storm surge of 2.4 m, and 193 mm of rainfall.[69] This unusual sequence of two major hurricanes created significant environmental challenges, as floodwaters and high winds from Hurricane Milton redistributed debris and contaminants initially displaced and temporarily stored by Hurricane Helene.[70–72]

We collected samples from 30 sites on 1 week, 2 weeks, and 7 months after Hurricane Milton. Sampling sites were classified as middle estuary if they lay along the river corridor and were bordered by land on both sides, and as lower estuary if they were located near the open coast where marine influence dominates. The sampling sites are shown in Fig. 1B, and the coordinates of the 30 sites are summarized in **Table S1**. The number of samples from Week 1, Week 2, and Month 7 are 30, 17, and 30 samples, respectively; thus, each site has two or three time points. There was one rainfall event observed at North Port Station in Florida between Week 2 and Month 7, with a recorded daily precipitation of 42 mm on February 25 in 2025; no other rainfall events exceeded 20 mm/day during this period (National Centers for Environmental Information). The 1- and 2-week post-storm samples represent near-term changes in water quality conditions after the two hurricanes, while the 7-month samples represent non-storm conditions.

The sampling was conducted by collecting 2 L of water (more than 30 cm below the surface) using a sterile HDPE bottle (2120-0005PK, Nalgene). All collected water samples were stored in coolers on ice and transported to the laboratory at the University of Florida within 3 days. New ice was added as needed to maintain sample temperatures. As quality assurance, travel blanks were prepared by adding DI water to sampling bottles in the laboratory and carried throughout the field trip.

Following the procedures described in Methods, we profiled bacterial community composition and quantified 12 indicators of contamination that can impair microbial water quality: terrestrial-associated pathogens (enterococci, *E. coli*, *Pseudomonas* spp.); marine-associated pathogens (*Vibrio parahaemolyticus*, *Vibrio vulnificus*); and six ARGs of public-health concern (*sul1*, *sul2*, *qnrS*, *tetA*, *floR*, and *aadA*).

4.2. Sample processing

Upon arrival at the laboratory, all water samples underwent a filtration process through 0.45 μm pore size sterile membrane filters with a diameter of 47 mm (HAWG047S6, MilliporeSigma) using a glass vacuum filtration system. The volume of filtered water varied, ranging from 0.12 to 0.70 L, depending on water

quality. One entire filter membrane from each site and time point was used to analyze the enterococci level following EPA method 1600. Specifically, the membrane containing the bacterial cells is placed on a 60 mm mEI agar plate (M13-140, Alpha Biosciences), prepared following the manufacturer's instructions. The bacteria were incubated 24 hours at 41°C ± 0.5°C. All colonies greater than or equal to 0.5 mm in diameter (regardless of color) with a blue halo were recorded as enterococci colonies. The filtration blanks were also prepared by passing DI water through all processes, from water filtration to the enterococci assay, and all results were negative.

Another set of filter membranes were cut into quarters. DNA was extracted from one or two quarters of each membrane using the FastDNA SPIN Kit for Soil (6560300, MP Biomedicals) according to the manufacturer's instructions, with a minor modification. Specifically, quarters were cut into pieces and placed into a Lysing Matrix E tube containing 978 µL of Sodium Phosphate Buffer and 122 µL of MT Buffer. The tube was loaded into TissueLyser II (QIAGEN) and homogenized with an intensity of 40 oscillations per second for 60 seconds. The final DNA was eluted in 60 µL of DES Elution Solution and stored at − 80°C until further analysis including DNA quantification and 16S rRNA amplicon sequencing.

4.3. DNA Quantification

Details about our dPCR analysis are described following the Minimum Information for Publication of Quantitative Digital PCR Experiments (dMIQE guidelines; **Table S2**). DNA obtained from the nucleic acid extraction was directly used for quantification using a dPCR system with the QIAcuity EvaGreen (EG) PCR Kit (250111, QIAGEN). Specifically, the dPCR mixture for nine analytes (total 16S rRNA, two pathogens including *ybbW* gene of *E. coli*, *orpI* gene of *Pseudomonas*), and six ARGs including Sul1, Sul2, qnrS, tetA, floR, and aadA) included 2 µL of nucleic acid extract, 0.6 µL of 8 µM forward primer, 0.6 µL of 8 µM reverse primer, 4 µL of 3x EvaGreen PCR Master Mix, and 4.8 µL of water. The other two analytes (*tlh* gene of *Vibrio parahaemolyticus* and *vvhA* gene of *Vibrio vulnificus*) were quantified with the QIAcuity Probe PCR kit (250101, QIAGEN). The dPCR mixture contained 2 µL of DNA, 0.4 µL of 24 µM forward primer, 0.4 µL of 24 µM reverse primer, 0.4 µL of 12 µM probe, 3 µL of 4x Probe PCR Master Mix, and 5.8 µL of RNase-free water.

Primer information for the 11 analytes quantified using dPCR is provided in **Table S3**. We confirmed that the primer sets synthesized a single and strong band with the expected amplicon size on gel electrophoresis (E-Gel™ Power Snap Electrophoresis System - G8100, Invitrogen). In addition, no primer dimer formation was confirmed (**Fig. S1**). This finding indicates that these primers specifically amplified the target sequences, ensuring that positive dPCR partitions are true positives and did not result from non-specific amplification.

The dPCR mixture was loaded onto an 8.5k 96-well nanoplate with a fixed partition volume of 0.34 nL and analyzed using the QIAcuity One 5-plex digital PCR system (QIAGEN). The thermal cycling protocol for the EG PCR kit consisted of one cycle at 95°C for 2 minutes, followed by 40 cycles of denaturation at 95°C for 15 seconds, annealing at 62°C for 15 seconds, and extension at 72°C for 15 seconds, with a final cooling step at 40°C for 5 minutes. The thermal cycling protocol for the Probe PCR kit included one

cycle at 95°C for 2 minutes, followed by 40 cycles of denaturation at 95°C for 15 seconds, and combined annealing and extension at 63°C for 30 seconds. The dPCR image was captured using the Green channel with an exposure duration of 200 ms and a gain of 6.

All partitions from dPCR analyses were separated into positive or negative with threshold for fluorescence intensity between 50 and 70 RFU with the QIAcuity EvaGreen (EG) PCR kit and between 10 and 15 RFU with the QIAcuity Probe PCR kit (**Fig. S2**). DNA was diluted in molecular biology-grade water using multiple dilution factors if necessary, and the measurement with the number of positive partitions between 1,000 and 5,000 was selected to determine the final DNA or RNA concentration when available. The limit of detection was 2 gene copies/µL when the number of positive partitions was 1, at which the 95% confidence interval (CI) was 147.5%. The number of valid partitions was determined automatically using information from the reference channel. We confirmed that the coefficient of variation in dPCR analysis was less than 5%, indicating that the repeatability of dPCR was acceptable (**Fig. S3**). As a result, one technical replicate was analyzed per sample. The confidence interval (CI) value determined by the number of positive and negative partitions for all analyses was less than 100%. All dPCR analyses were conducted with molecular-grade water as a negative control. The concentration of the negative control was subtracted from each measurement if a negative control exhibited positive partitions. We conducted an inhibition test for the undiluted nucleic acid extracts, and the impact of inhibitors in the extracts was negligible for dPCR analysis (**Fig. S3**).

The travel blanks went through the same processes, including sampling campaign, water filtration, DNA extraction process, and dPCR analysis. These samples tested negatives for all analytes, except for 16S rRNA. Although we autoclaved all possible materials and conducted DNA extraction and dPCR preparation inside biosafety cabinets (certified following the NSF/ANSI 49 standard), 16S rRNA was still detected in the travel and PCR blanks (which is less than 1% of 16S rRNA concentration in water samples). We subtracted these ambient 16S rRNA signals from the water samples.

4.4. 16S rRNA amplicon sequencing

The bacterial community of water samples was investigated using 16S rRNA amplicon sequencing. We amplified about 1,500 bp-long bacterial 16S rRNA gene using a PCR thermocycler (Mastercycler x40, Eppendorf). The primer information is provided in **Table S3**. The thermal cycling protocol consisted of one cycle at 94°C for 2 minutes, followed by 35 cycles of denaturation at 94°C for 15 seconds, annealing at 53°C for 15 seconds, and extension at 68°C for 30 seconds, with a final holding step at 4°C. Gel electrophoresis confirmed single, clear 1.5 kb bands, indicating successful amplification of bacterial 16S rRNA (**Fig. S1**).

The PCR amplicons were purified using a Quick PCR purification kit (K310001, Invitrogen). This process satisfied all requirements regarding sequencing quality control suggested by Oxford Nanopore Technologies (ONT, the United Kingdom). For example, we quantified DNA concentrations of PCR amplicons using a fluorometer (Qubit 4, Invitrogen) with a Qubit 1X dsDNA HS Assay Kit (Q33231, Thermo Scientific). All input DNA concentrations were higher than 200 fmol (i.e., 200 ng for 1.5 kb

amplicons). We also checked the DNA purity using a spectrophotometer (Nanodrop Ultra FL, Thermo Fisher Scientific), and all samples had a 260/280 ratio of approximately 1.80 and a 260/230 ratio between 2.0 and 2.2, indicating minimal impact of protein, RNA, phenol, or other contaminants on sequencing.

The purified amplicons were used for library preparation with the SQK-NBD114-96 kit (ONT, the United Kingdom). Barcoded DNA samples were pooled and loaded onto a flow cell (FLO-MIN114, Oxford Nanopore Technologies). Sequencing was conducted with MinION (Oxford Nanopore Technologies) for 3 days. We ran the super-accurate basecalling model (v5.0.0) using a supercomputing system with 4 CPUs and 1 GPU (i.e., HiperGator at the University of Florida) to process raw POD5 files generated by the sequencing system. The raw sequencing data is available in the NCBI under BioProject (PRJNA1363861). Sequenced reads were filtered using a Q score threshold greater than 20 (i.e., 99% base accuracy) and a minimum read length of 100 bp to ensure species-level classification in microbial community analysis. As a result, an average of 23,423 reads per sample was obtained, totaling about 1.8 million reads across all samples (**Table S4**). Taxonomic annotation was performed at the species level using an ONT workflow of 16S rRNA analysis with the Kraken2 classifier.

4.5. Geochemical parameters

Water samples were also collected at the same sites to evaluate CDOM composition along with carbon and nitrogen stable isotopes in particulate material as an independent measure of terrestrial and marine organic matter. CDOM samples were filtered through 0.2 µm PES filters and particulate material was collected on pre-combusted glass fiber filters (GFFs). CDOM samples were analyzed using a spectrofluorometer (Horiba Aqualog®) with excitation wavelengths ranging from 240 to 800 nm (3 nm increments) and emission wavelengths ranging from 246 to 827 nm (2.3 nm increments) and data were analyzed in R using the StaRdom package[73]. The absorption coefficient at 254 nm (a254) was used as a proxy of the optically active dissolved organic matter pool (CDOM). Particulate material was freeze dried, rolled into tin capsules and analyzed for carbon ($\delta^{13}C$) and nitrogen stable isotopes ($\delta^{15}N$) using a Carlo Erba 1500 CN elemental analyzer coupled to a Thermo Electron Delta Advantage isotope ratio mass spectrometer (Carlo Erba/ThermoFisher ScientificTM, Waltham, MA, USA). Salinity was measured *in situ* using handheld multiparameter probes.

4.6. Large-Language-Model-Assisted Microbial Source Tracking (LAMST) development

We developed a novel Large-Language-Model-Assisted Microbial Source Tracking (LAMST) framework to trace the sources of microbial contamination in the samples (Fig. 2A**)**. We employed Kraken2 to identify the bacterial species and its abundance in each sample (**Step 1 in** Fig. 2A). The number of species identified in each sample ranged from 41 to 407, with a total of 6,172 bacterial species identified across all samples. We fetched metadata for each microbial species in a XML format from NCBI Biosample database using NCBI Entrez Direct with *esearch* and *efetch* commands (**Step 2 in** Fig. 2A). The number of metadata records varied substantially depending on bacterial species. For example, *Staphylococcus*

aureus, Klebsiella pneumoniae, and *Pseudomonas aeruginosa* are bacterial species showing the largest metadata upload at NCBI BioSample database (198194, 167675, and 98523 records, respectively), while 1,090 bacterial species had no associated metadata. The bacterial species lacking metadata were labeled Insufficient and excluded from further analysis. We compiled ~ 13 million metadata records across all species with 3482 unique attributes. The five most frequent attributes were: collection_date (669,118 entries), geo_loc_name (660,585), isolation_source (633,072), strain (567,846), and host (564,004). While some attributes—such as isolation_source and host—capture sample origin, they are recorded as unstructured text. For example, entries for isolation_source include “AGS wastewater treatment plant,” “River water sample from the Mahananda River near Suryasen Park,” and “Produced fluids from hydraulically fractured shales.” Across ~ 13 million text entries spanning 3,482 attributes, this unstructured format is difficult to organize for inferring microbial origins. This challenge motivated us to use LLMs to interpret unstructured data.

We extracted answers for seven attributes: isolation source, host, broad-scale environmental context, environmental medium, sample type, metagenomic source, local environmental context, from XML files (**Step 3 in** Fig. 2A). These answers were used as input for Llama 3-8B (Meta Platforms. Inc., USA). We focused on these fields rather than using the entire XML files because Llama has a maximum input length of 8000 tokens (~ 20,000 words), while some XML files exceeded this limit. For cases where the extracted key words exceeded the token limit, we retained the most recent entries up to 7900 tokens. Each species-level metadata was provided to Llama with the following prompt:

“*You are an environmental ecologist. Each keyword listed below (separated by semicolons) represents a value related to one of the following categories: isolation source, host, broad-scale environmental context, environmental medium, sample type, metagenomic source, or local environmental context. Based solely on this information (without relying on any prior knowledge), if this microbial species is detected in an environmental sample, what is the most likely environment where it is found? If any keywords are related to marine animals, consider them as indicators of a marine environment. If any keywords are related to humans, sludge, WWTP or hospitals, consider them as indicators of a wastewater environment. Choose only one of the following: Marine, Terrestrial, Wastewater, or Insufficient. Only respond with the one-word answer after 'Answer:' and nothing else*”.

After repeating this query to all microbial species in a sample (**Step 4 in** Fig. 2A), we summed the abundances of each category: Marine, Terrestrial, Wastewater, and Insufficient. The relative abundance of Marine, Terrestrial, and wastewater was computed by normalizing with the total abundance of Marine, Terrestrial, and Wastewater bacteria, while excluding the Insufficient class from the normalization (**Step 5 in** Fig. 2A). This process provided the relative abundance of Marine, Terrestrial, and Wastewater bacteria at each site. The script for running LAMST is available at https://github.com/ChamteutOh/LAMST.

4.7. Large language model-assisted microbial source tracking (LAMST) validation

To validate LAMST, we first applied it to sequences collected from marine, terrestrial, and wastewater samples (Fig. 2B). The sequencing data was downloaded from NCBI Sequence Read Archive (SRA). We used two search key words: either *marine sediment*, *soil*, or *wastewater* and *16S* to fetch sequences from marine, terrestrial, and wastewater samples, respectively. The search results were screened by two filters: *DNA* and *Oxford Nanopore*. Each SRA deposit has an abstract about the samples and analytical methodology. We manually downloaded sequences obtained with a long-range 16S rRNA amplicon sequencing from the SRA database, as the length of 16S rRNA amplicon is the key to annotating microbial taxa on a species level. We downloaded 20 marine samples, 16 terrestrial samples, and 23 wastewater samples. These reference sequences were processed with LAMST to assess how inferred bacterial habitats vary across sampling environments.

4.8. Statistical analysis

We compared analyte levels at each site between Week 1 and Month 7 to assess whether Hurricane Milton affected microbial water quality, and between Week 1 and Week 2 to evaluate short-term changes (Fig. 1C-P). A Shapiro–Wilk test was applied to the paired differences between groups. If the differences were normally distributed ($p \geq 0.05$), we used a paired t-test; otherwise ($p < 0.05$), we used a Wilcoxon signed-rank test. The selected test, its p-value, and the corresponding Shapiro–Wilk p-value are reported in **Table S4**.

For Fig. 2, we compared two independent groups. We first applied a Shapiro–Wilk test to each group to assess normality. If both groups were approximately normal ($p \geq 0.05$), we used Welch's t-test; otherwise, we used the Mann–Whitney U test. Detailed information on the statistical analyses is summarized in **Table S5**.

For Fig. 3, we quantified how microbial-origin abundances relate to microbial-threat concentrations (FIB, pathogens, and ARGs) using site fixed-effects (FE) regressions. To assess the effect of transforming predictors, each predictor–analyte pair was fit twice—once with raw abundances and once with log1p-transformed abundances. Across 225 pairs, the log1p models yielded a lower median p-value than the raw models (0.170 vs 0.289), indicating stronger overall associations. Accordingly, we report and interpret the log1p-based results for evaluating links between microbial-origin abundances and microbial-threat concentrations.

Relationships between the relative abundance of terrestrial, marine, and wastewater bacteria and geochemical measures (CDOM, stable isotopes) were also evaluated to further validate LAMST. Pearson correlations were run in R version 4.3.0 using the package ggpubr[74].

Declarations

Data Availability

All sequencing data has been submitted to the NCBI BioProject (PRJNA1363861) and the code is available at https://github.com/ChamteutOh/LAMST.

Acknowledgement

This study was funded by a National Science Foundation RAPID award to CO, SJ, and ESM (Award Numbers: CBET 2519559) and the National Academies Early Career Research Fellowship to ESM. The authors would like to thank Christine Angelini, Todd Van Natta, Patrick Norby, Hallie Fischman, and Zanti Rains for their assistance in the field, and AJ Reisinger and Jason Curtis for laboratory assistance.

References

1. Bianucci L, Balaguru K, Smith RW, Leung LR, Moriarty JM (2018) Contribution of hurricane-induced sediment resuspension to coastal oxygen dynamics. Sci Rep 8:1–10
2. Tian M et al (2024) Sediment resuspension and transport in the offshore subaqueous Yangtze Delta during winter storms. Front Mar Sci 11:1420559
3. Curran JF, Zaggia L, Quero GM (2022) Metagenomic Characterization of Microbial Pollutants and Antibiotic-and Metal-Resistance Genes in Sediments from the Canals of Venice. Water (Switzerland) 14:1161
4. Hatosy SM, Martiny AC (2015) The Ocean as a Global Reservoir of Antibiotic Resistance Genes. Appl Environ Microbiol 81:7593–7599
5. Li Z et al (2024) Metagenomic analysis manifesting intrinsic relatedness between antibiotic resistance genes and sulfate- and iron-reducing microbes in sediment cores of the Pearl River Estuary. Environ Pollut 363:125176
6. Jia L et al (2022) Distribution and Transfer of Antibiotic Resistance Genes in Coastal Aquatic Ecosystems of Bohai Bay. Water (Switzerland) 14:938
7. Zhao Z et al (2020) Estuarine sediments are key hotspots of intracellular and extracellular antibiotic resistance genes: A high-throughput analysis in Haihe Estuary in China. Environ Int 135:105385
8. Magalhães EA, de Jesus HE, Pereira PHF, Gomes AS, Santos HF (2024) dos. Beach sand plastispheres are hotspots for antibiotic resistance genes and potentially pathogenic bacteria even in beaches with good water quality. Environ Pollut 344:123237
9. Johnson LA et al (2025) Influence of aquaculture and genomic surveillance techniques on antimicrobial resistance gene profiles and microbiota detected in marine and freshwater sediments. https://doi.org/10.1139/cjm-2024-0206 https://doi.org/10.1139/CJM-2024-0206 doi:10.1139/CJM-2024-0206
10. Moritz CF, Snyder RE, Riley LW, Immke DW, Greenfield BK (2020) Antimicrobial Drug-Resistant Gram-Negative Saprophytic Bacteria Isolated from Ambient, Near-Shore Sediments of an Urbanized Estuary: Absence of β-Lactamase Drug-Resistance Genes. Antibiot 2020 9:400
11. Chen Y, ru et al (2020) Antibiotic resistance genes (ARGs) and their associated environmental factors in the Yangtze Estuary, China: From inlet to outlet. Mar Pollut Bull 158:111360
12. Phillips MC et al (2014) Microbial release from seeded beach sediments during wave conditions. Mar Pollut Bull 79:114–122

13. Levin PS, Howe ER, Robertson JC (2020) Impacts of stormwater on coastal ecosystems: the need to match the scales of management objectives and solutions. Philosophical Trans Royal Soc B 375
14. Yu P et al (2018) Elevated Levels of Pathogenic Indicator Bacteria and Antibiotic Resistance Genes after Hurricane Harvey's Flooding in Houston. Environ Sci Technol Lett 5:481–486
15. Yang SH, Chen CH, Chu KH (2021) Fecal indicators, pathogens, antibiotic resistance genes, and ecotoxicity in Galveston Bay after Hurricane Harvey. J Hazard Mater 411:124953
16. Ahmed W et al (2018) Precipitation influences pathogenic bacteria and antibiotic resistance gene abundance in storm drain outfalls in coastal sub-tropical waters. Environ Int 116:308–318
17. Hou L et al (2022) Storm promotes the dissemination of antibiotic resistome in an urban lagoon through enhancing bio-interactions. Environ Int 168:107457
18. Han Y et al (2022) Occurrence and driving mechanism of antibiotic resistance genes in marine recreational water around Qinhuangdao, China. Front Mar Sci 9:976438
19. Cosby AG et al (2024) Accelerating growth of human coastal populations at the global and continent levels: 2000–2018. Sci Rep 14:1–10
20. Levy K, Hubbard AE, Nelson KL, Eisenberg J (2009) N. S. Drivers of Water Quality Variability in Northern Coastal Ecuador. Environ Sci Technol 43:1788–1797
21. Paruch AM, Paruch L (2024) Current status of microbial source tracking applications in constructed wetlands serving as nature-based solutions for water management and wastewater treatment. Environ Pollut 351:124076
22. Wang XW et al (2023) Ecological dynamics imposes fundamental challenges in community-based microbial source tracking. iMeta 2:e75
23. Mayer RE et al (2018) Global Distribution of Human-Associated Fecal Genetic Markers in Reference Samples from Six Continents. Environ Sci Technol 52:5076–5084
24. Hill ER, Chun CL, Hamilton K, Ishii S (2023) High-Throughput Microfluidic Quantitative PCR Platform for the Simultaneous Quantification of Pathogens, Fecal Indicator Bacteria, and Microbial Source Tracking Markers. ACS ES&T Water 3:2647–2658
25. De La Barron C (2023) Quantification of the mobility potential of antibiotic resistance genes through multiplexed ddPCR linkage analysis. FEMS Microbiol Ecol 99:1–10
26. Schoonbroodt S et al (2022) Real-time PCR has advantages over culture-based methods in identifying major airway bacterial pathogens in chronic obstructive pulmonary disease: Results from three clinical studies in Europe and North America. Front Microbiol 13:1098133
27. Gezer U, Bronkhorst AJ, Holdenrieder S (2022) The Clinical Utility of Droplet Digital PCR for Profiling Circulating Tumor DNA in Breast Cancer Patients. Diagnostics 2022 12(12):3042
28. Tambalo DD, Fremaux B, Boa T, Yost CK (2012) Persistence of host-associated Bacteroidales gene markers and their quantitative detection in an urban and agricultural mixed prairie watershed. Water Res 46:2891–2904

29. Lobos AE et al (2024) Persistence of sewage-associated genetic markers in advanced and conventional treated recycled water: implications for microbial source tracking in surface waters. *mBio* 15
30. Knights D et al (2011) Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8:761–765
31. Zhang W et al (2023) Comprehensive Assessment of 16S rRNA Gene Amplicon Sequencing for Microbiome Profiling across Multiple Habitats. Microbiol Spectr 11
32. Segata N et al (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814
33. Huang Z, Cai D, Sun Y (2024) Towards more accurate microbial source tracking via non-negative matrix factorization (NMF). Bioinformatics 40:i68–i78
34. Bowen M, Farag IF, Main CR, Biddle JF, Dennehy JJ (2024) Reference library for microbial source tracking in the mid-Atlantic United States. Microbiol Resour Announc 13
35. Staley C et al (2018) Application of SourceTracker for Accurate Identification of Fecal Pollution in Recreational Freshwater: A Double-Blinded Study. Environ Sci Technol 52:4207–4217
36. LaMontagne MG, Zhang Y, Guillen GJ, Gentry TJ, Allen MS (2022) Hurricane Harvey Impacts on Water Quality and Microbial Communities in Houston, TX Waterbodies. Front Microbiol 13:875234
37. Wang Z et al (2025) Accelerating clinical evidence synthesis with large language models. NPJ Digit Med 8:1–14
38. Zhang Y et al (2025) Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence 2025 1:1* 1, 1–15
39. Luo X et al (2024) Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. J Med Internet Res 26
40. Polak MP, Morgan D (2024) Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 15:1–11
41. Gougherty AV, Clipp HL (2024) Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. NPJ Biodivers 3:1–5
42. Florida Healthy Beaches | Florida Department of Health https://www.floridahealth.gov/environmental-health/beach-water-quality/index.html
43. Tillman MC et al (2025) Acute Impacts of Hurricane Ian on Benthic Habitats, Water Quality, and Microbial Community Composition on the Southwest Florida Shelf. Coasts 2025 5(5):16
44. Warnakulasooriya KN, Butler ECV, Gibb KS, Munksgaard NC (2017) Stable isotopes in biota reflect the graduated influence of sewage effluent along a tropical macro-tidal creek. Mar Freshw Res 68:1855–1866
45. Cole ML et al (2004) Assessment of a δ 15 N Isotopic Method to Indicate Anthropogenic Eutrophication in Aquatic Ecosystems. J Environ Qual 33:124–132

46. Huang C et al (2020) Carbon and nitrogen isotopic composition of particulate organic matter in the Pearl River Estuary and the adjacent shelf. Estuar Coast Shelf Sci 246:107003
47. Florida Cattle Market Update – January 2025 Annual Cattle Market Report | Panhandle Agriculture. https://nwdistrict.ifas.ufl.edu/phag/2025/02/07/florida-cattle-market-update-annual-cattle-market-report-january-2025/
48. (2023) Summary Report On Antimicrobials Sold or Distributed for Use in Food-Producing Animals | FDA. https://www.fda.gov/animal-veterinary/antimicrobial-resistance/2023-summary-report-antimicrobials-sold-or-distributed-use-food-producing-animals
49. USDA - National Agricultural Statistics Service - Florida - County Estimates. https://www.nass.usda.gov/Statistics_by_State/Florida/Publications/County_Estimates/index.php
50. Letchumanan V, Chan KG, Lee LH (2014) Vibrio parahaemolyticus: a review on the pathogenesis, prevalence, and advance molecular identification techniques. Front Microbiol 5:705
51. Lopatek M, Wieczorek K, Osek J (2015) Prevalence and Antimicrobial Resistance of Vibrio parahaemolyticus Isolated from Raw Shellfish in Poland. J Food Prot 78:1029–1033
52. Osborne B, Siboni N, Seymour JR, Ralph P, Pernice M (2023) Exploring the potential of algae-bacteria interactions in the biocontrol of the marine pathogen Vibrio parahaemolyticus. J Appl Phycol 35:2731–2743
53. Consequences of Hurricanes Helene and Milton in Florida https://floridaspecifier.com/issues/v46n6/assessing-the-environmental-consequences-of-hurricanes-helene-and-milton-in-florida/
54. Pollution spills caused by hurricane season https://environmentamerica.org/florida/center/resources/pollution-spills-caused-by-hurricane-milton/
55. Hurricane Milton | Disasters, Storms, & Florida | Britannica. https://www.britannica.com/event/Hurricane-Milton
56. Sowah RA, Molina M, Georgacopoulos O, Snyder B, Cyterski M (2022) Sources and Drivers of ARGs in Urban Streams in Atlanta, Georgia, USA. *Microorganisms* 10, 1804
57. Su Z et al (2020) Wastewater discharge drives ARGs spread in the coastal area: A case study in Hangzhou Bay, China. Mar Pollut Bull 151:110856
58. O'Malley K, McDonald W, McNamara P (2023) Antibiotic resistance in urban stormwater: a review of the dissemination of resistance elements, their impact, and management opportunities. Environ Sci (Camb) 9:2188–2212
59. Barancheshme F, Munir M (2018) Strategies to combat antibiotic resistance in the wastewater treatment plants. Front Microbiol 8:299736
60. Londong J, Barth M, Söbke H (2023) Reducing antimicrobial resistances by source separation of domestic wastewater. Front Environ Health 2

61. Contti Neto N, Pomeroy A, Lowe R, Ghisalberti M (2022) Seagrass Meadows Reduce Wind-Wave Driven Sediment Resuspension in a Sheltered Environment. Front Mar Sci 8:733542
62. Tokuda M, Shintani M (2024) Microbial evolution through horizontal gene transfer by mobile genetic elements. Microb Biotechnol 17:e14408
63. Johansson MHK, Aarestrup FM, Petersen TN (2023) Importance of mobile genetic elements for dissemination of antimicrobial resistance in metagenomic sewage samples across the world. PLoS ONE 18:e0293169
64. Jampani M et al (2024) Fate and transport modelling for evaluating antibiotic resistance in aquatic environments: Current knowledge and research priorities. J Hazard Mater 461:132527
65. Peng J et al (2024) Seasonal dynamics of antibiotic resistance genes and mobile genetic elements in a subtropical coastal ecosystem: Implications for environmental health risks. Environ Res 257:119298
66. Yamahara KM, Sassoubre LM, Goodwin KD, Boehm AB (2012) Occurrence and Persistence of Bacterial Pathogens and Indicator Organisms in Beach Sand along the California Coast. Appl Environ Microbiol 78:1733
67. Frank EM et al (2025) Microbial Contamination in Urban Marine Sediments: Source Identification Using Microbial Community Analysis and Fecal Indicator Bacteria. Microorganisms 13:983
68. Lu X, Yu S, Zhang L, Liu S, Lu H (2025) Microbial signatures and ARGs profiles in two coastal ecosystems of Shenzhen under distinct anthropogenic influences. iScience 28:113133
69. Hurricanes | Sarasota Bay Estuary Program https://sarasotabay.org/our-estuaries/climate-change/hurricanes/
70. ‘It’s a whole different level of destruction:’ St. Armands Circle lined with Hurricane Helene damage. https://www.mysuncoast.com/2024/10/01/its-whole-different-level-destruction-st-armands-circle-lined-with-hurricane-helene-damage/?utm_source=chatgpt.com
71. Florida digs out of mountains of sand swept in by back-to-back hurricanes | AP News. https://apnews.com/article/milton-helene-hurricane-sand-florida-2906dfdc1027051547c4d49e934a838d
72. Sarasota faces months of debris collection | Your Observer https://www.yourobserver.com/news/2024/oct/21/city-debris-collection/?utm_source=chatgpt.com
73. Pucher M et al (2019) staRdom: Versatile Software for Analyzing Spectroscopic Data of Dissolved Organic Matter in R. Water 2019 11(11):2366
74. Ggpubr ‘ggplot2’ based publication ready plots | CiNii Research. https://cir.nii.ac.jp/crid/1370861707141778052

Figures

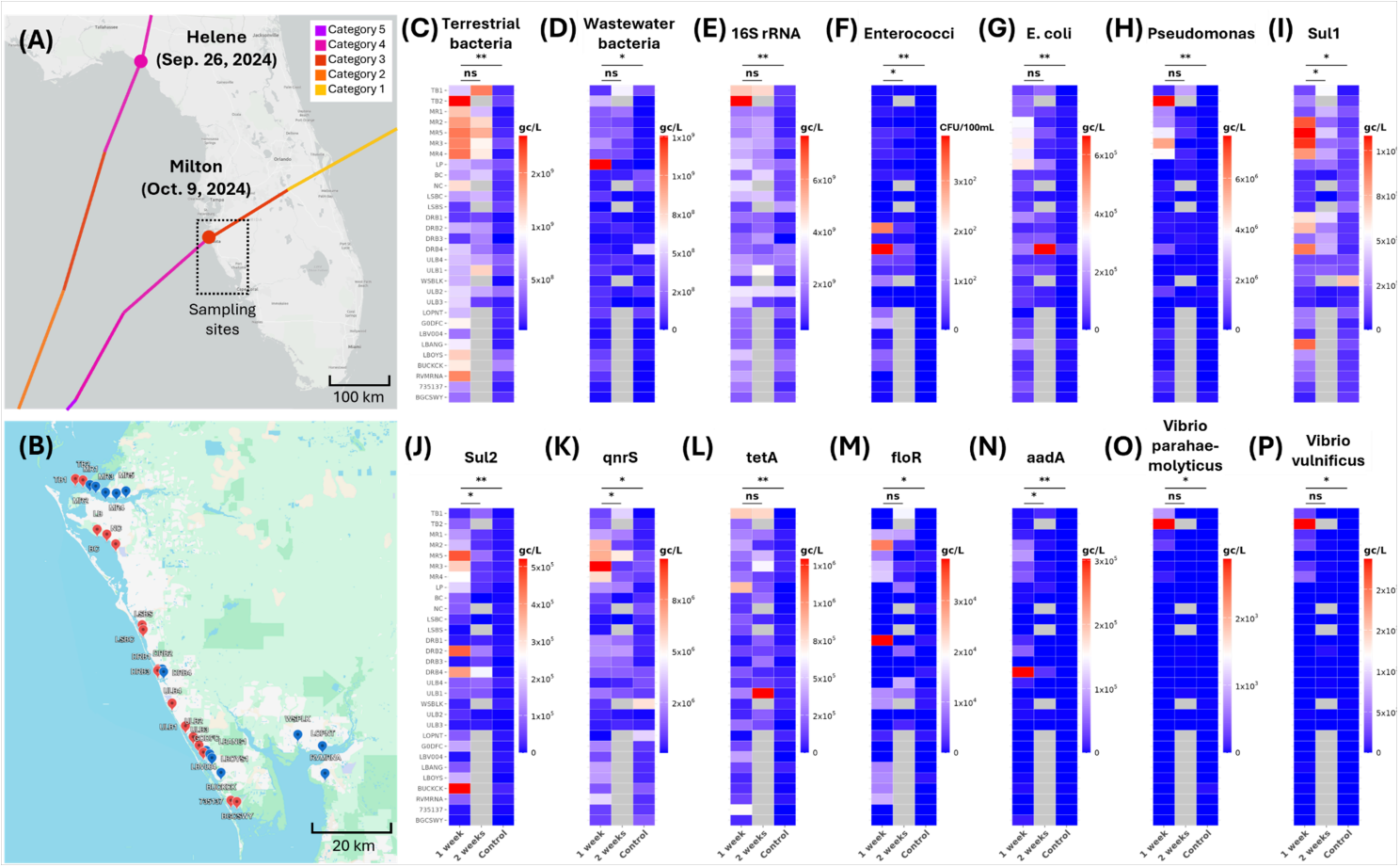

Figure 1

(A) Trajectories of Hurricanes Helene and Milton with temporal changes in hurricane intensity (https://coast.noaa.gov/hurricanes/), (B) Locations of the thirty sampling sites; blue and red arrows indicate the middle and lower estuaries, respectively. (C–P) Heat maps showing the relative abundance of terrestrial and wastewater bacteria (C and D, respectively), and water quality measurements by CFU assay (F) and dPCR (other panels). Rows of the heat maps indicate sampling sites, organized from highest to lowest latitude. Columns represent sampling times of 1 week, 2 weeks, and control (7 months). Red color indicates higher values, blue indicates lower values, and gray cells indicate missing data. Full statistical results are provided in **Table S4**.

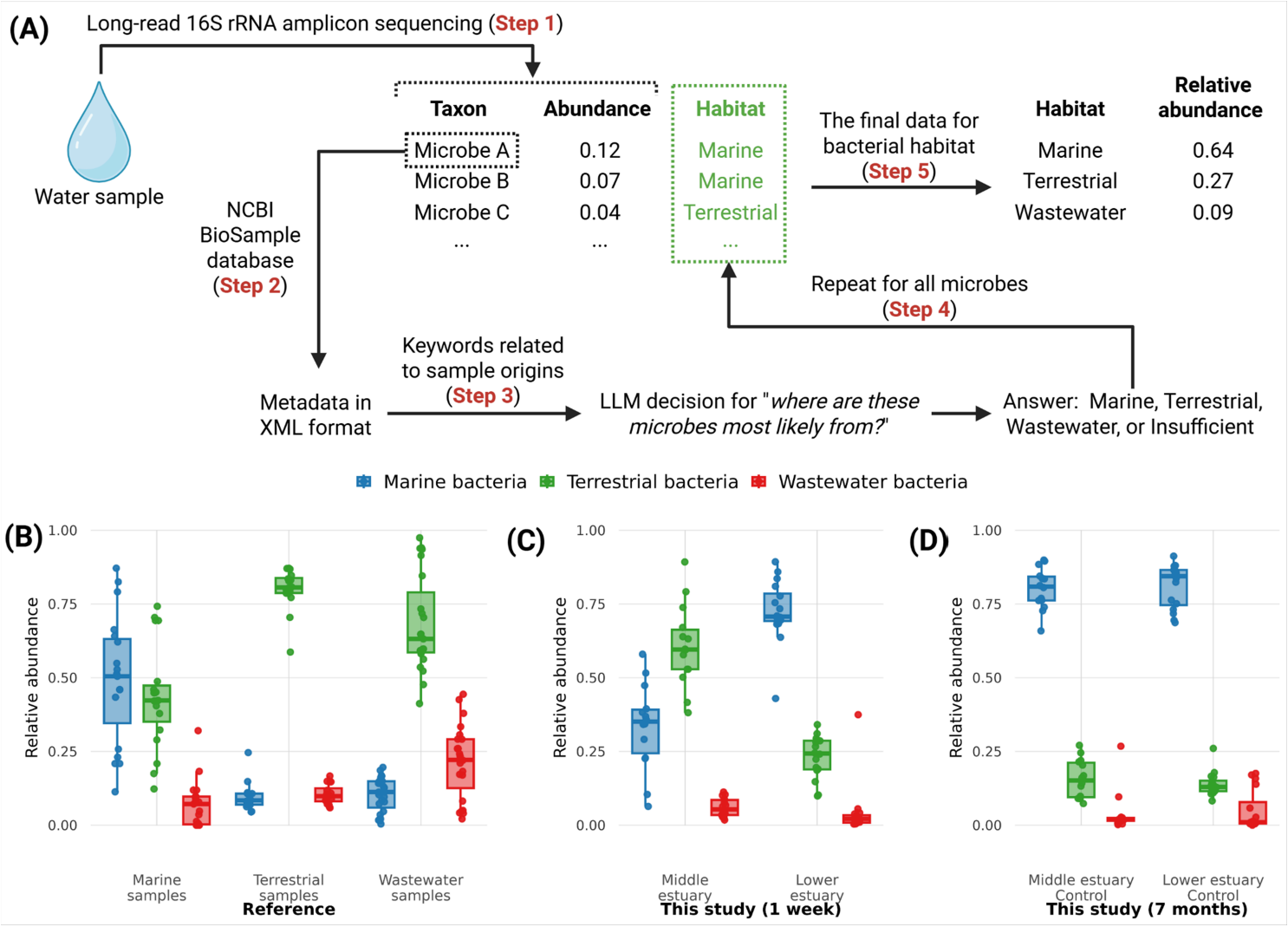

Figure 2

(A) Schematic of the Large-Language-Model-Assisted Microbial Source Tracking (LAMST) framework for determining microbial origin in water samples. (B) Validation of the framework using reference sequences deposited in the NCBI Sequence Read Archive (SRA) from marine, terrestrial, and wastewater samples. (C) Application of LAMST to water samples collected one week after Hurricane Milton from the middle and lower estuary, and to (D) samples collected from the same sites seven months after Hurricane Milton. Full statistical results are provided in **Table S5**.

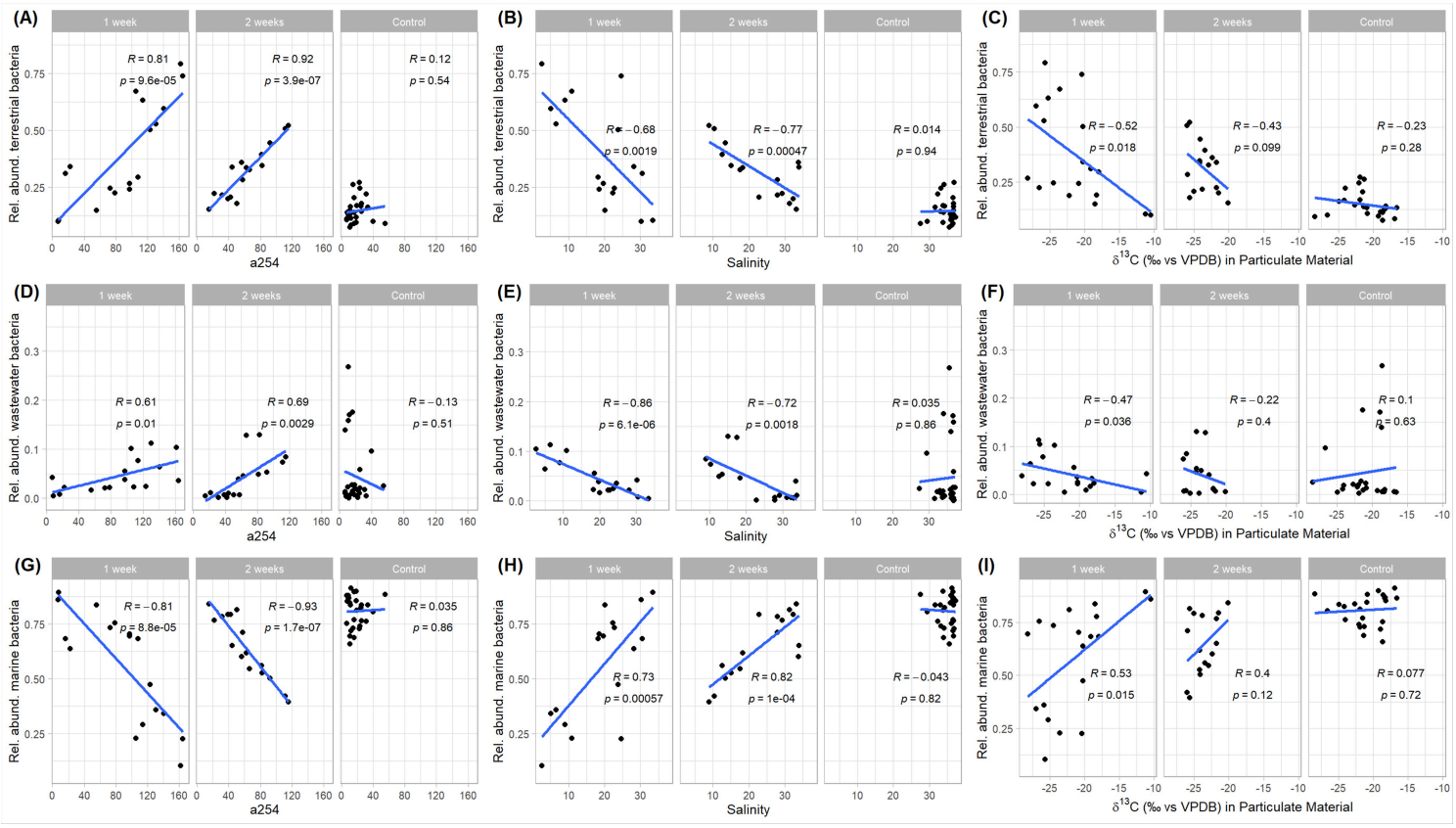

Figure 3

Relationships between the relative abundance of terrestrial bacteria (upper panels; A, B, and C), wastewater bacteria (D, E, and F), and marine bacteria (G, H, and I) compared with colored dissolved organic matter (a254; left panels), salinity (middle panels), and carbon stable isotopes in particulate material (right panels). *R* and *p* values are shown for Pearson correlations.

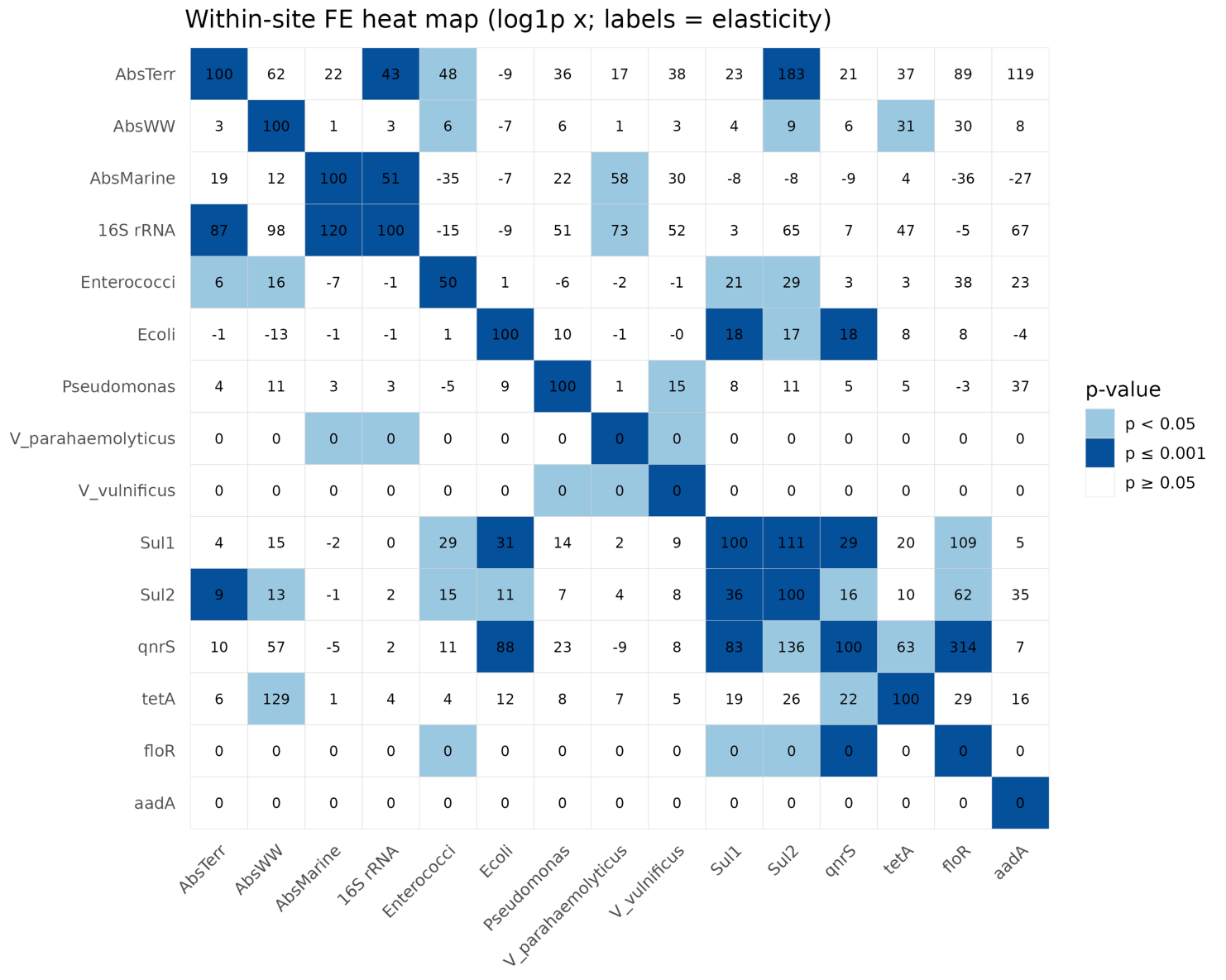

Figure 4

Heat map of site fixed-effects (FE) regressions using log1p-transformed data for fourteen variables. Each cell shows the elasticity—the approximate percent change in the analyte for a doubling of the predictor (positive = increase; negative = decrease). Colors indicate the within-site permutation p-values. Note that zero elasticities for *V. parahaemolyticus*, *V. vulnificus*, floR, and aadA reflect the presence of zero measurements, which leave minimal within-site variation.

Supplementary Files

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

- MiltonwaterSI6.docx
- floatimage1.png