Chesapeake Community Research Symposium 2026

Session 19: Advanced Data Analytics for Water Quality and Public Health

Session Leads: Jianyong Wu, Dongmei Alvi, & Efeturi Oghenekaro

---

Emily H. Majcher (U.S. Geological Survey MD-DE-DC Water Science Center)

Status of PFAS in surface waters of the Chesapeake Bay Watershed and associations with sources and landscape characteristics

Per and polyfluoroalkyl substances (PFAS), a group of chemical compounds that have been used in a variety of industries in the United States since the 1940s, are widespread and persistent in the environment. PFAS pose risks to human and ecosystem health, even at very low concentrations. Consistent monitoring to foster assessments of contaminants of emerging concern, such as PFAS, was prioritized within the Chesapeake Bay Program as part of recommendations for enhanced monitoring to the Principals' Staff Committee (2021) and as part of a Scientific and Technical Advisory Committee, state of the science workshop (2022). From 2023-2024, a synthesis of existing, public PFAS data from state and local jurisdictions, federal agencies, and academic institutions was conducted and these data were evaluated for their consistency and quality. This spatially extensive data set of PFAS in surface water will be described, highlighting detections, concentrations, and composition of PFAS in non-tidal and tidal surface waters of the watershed. Statistical and geographic analyses were used to examine the correlation between local and upstream sources and landscape characteristics with PFAS concentrations in non-tidal surface waters; the main goal was to identify key drivers of PFAS presence and absence in surface waters throughout the Chesapeake Bay watershed. Additionally, a potential atmospheric deposition fingerprint and potential background concentrations of PFAS in the Chesapeake Bay watershed were assessed and compared to other regions of the country. The PFAS data associated with this assessment will be included and updated in Chesapeake Data to allow for continued, ongoing interpretation.

Yanni Cao (Division of Environmental Health Sciences, College of Public Health, The Ohio State University)

Environmental Indicators of PFAS in an Urban Watershed Revealed by Explainable Machine Learning

Per- and polyfluoroalkyl substances (PFAS) are persistent "forever chemicals" that pose significant challenges to urban source-water protection. While industrial sources are often the focus, there is a critical gap in understanding how hydrometeorological factors and land use

interact to regulate PFAS fluxes to reservoirs. This study examined an urban watershed (1,480 km²) that serves approximately 1.0 million residents in the Chesapeake watershed. Monthly water samples were collected from January 2024 to October 2025 and analyzed for 40 PFAS compounds using LC–MS/MS following EPA Method 1633. PFAS concentrations below detection limits were estimated using a Bayesian approach. Water-quality parameters, weather data, and land-cover data were integrated with PFAS data. XGBoost was used as the optimal machine-learning model and SHAP (SHapley Additive exPlanations) quantified the nonlinear contributions of predictors. Total PFAS concentrations in the watershed averaged 228.0 ng/L (range: 13.2–5527.0 ng/L), with total PFCAs contributing the highest mean concentration at 135.1 ng/L. XGBoost demonstrated robust performance for total PFAS ($R^2 = 0.708$) and total PFCAs ($R^2 = 0.646$). SHAP analysis identified conductivity, developed land, total alkalinity, and total suspended solids as the primary predictors of PFAS levels, indicating that mobilization is strongly driven by anthropogenic land use and geochemical markers of urban runoff. By leveraging explainable AI, this research identifies key environmental indicators governing PFAS transport, providing actionable insights for adaptive water-quality monitoring and management.

Benjamin Schelling (Old Dominion University), Margaret Mulholland

The Effects of Tidal Flooding on the Transport of Per- & Polyfluoroalkyl Substances (PFAS) into the Lafayette and Elizabeth Rivers

Per- & polyfluoroalkyl substances (PFAS), a group of synthetic compounds, reside in organisms longer than any known anthropogenic contaminant, and the largest risk for human consumption of PFAS is through drinking water and seafood. PFAS are found in numerous man-made products including aqueous fire-fighting foam, cleaning products, and pesticides. Like many other organic contaminants, PFAS compounds have the potential to bioaccumulate in aquatic organisms and biomagnify in higher trophic levels. Understanding the pathways that PFAS compounds enter waterways is of high importance. One potential pathway is through runoff and tidal flooding, which carry nutrients and contaminants into coastal waterways. Tidal flooding has increased as a result of sea level rise. The Hampton Roads region of Virginia has one of the highest rates of sea level rise in the world. Tidal flooding is now routine during the spring tide portion of the tidal cycle in southeastern Virginia. To determine whether tidal flooding contributes to PFAS loading, a study was conducted in the Elizabeth and Lafayette Rivers, tidal tributaries of the James River and lower Chesapeake Bay. Water samples were collected from the estuary before, during, and after king tide events in Fall 2024. Floodwater samples were collected during high tide on September 21st, 2024, and October 19th, 2024, in conjunction with large citizen science campaigns (Measure the Muck). Results from water samples collected before and after periods of tidal flooding, and concentrations in floodwaters, were compared to estimate PFAS loads into this tidal estuary.

Dongmei Alvi (Occoquan Watershed Monitoring Laboratory)

Explainable AI Illuminates Microbial Contributions in a Highly Urbanized Watershed

Microbial contamination in urban watersheds presents persistent challenges for water quality management and public health. Traditional methods of analyzing monitoring results provide limited information about the sources of contamination and the environmental factors driving it. This study utilized explainable machine learning techniques to analyze microbial source data obtained from a chip-based digital PCR (cdPCR) platform. The goal was to identify the primary sources of bacterial contamination and the key environmental predictors in a highly urbanized waterway in Washington, DC. Biweekly water samples were collected over 13 months (2021–2022) at three representative sites. The modeling dataset included cultural fecal indicator bacteria (FIB) Escherichia coli (E. coli), four host-associated genetic markers (HF183, Rum2Bac, DG3, and GFD), in situ hydrochemophysical measurements, as well as meteorological and land-cover variables. Random forest regression and redundancy analysis (RDA) were employed to assess multivariate relationships. Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) quantified the contributions of various variables. HF183 showed the highest concentrations and frequent positive cdPCR detections, but SHAP consistently identified the canine-associated DG3 marker as the strongest predictor of microbial contamination. The model highlighted key environmental factors affecting variability, including storm events, minimum vapor pressure deficit, turbidity, and streamflow. This integrated framework facilitated diagnosis of microbial contamination dynamics by identifying key sources and their environmental controls, supporting targeted water quality management strategies.

Anna Van Dongen (Smithsonian Environmental Research Center)

Public Health and Water Quality: Investigating the Relationship between Bacteria and Colored Dissolved Organic Matter

Chesapeake Bay is home to millions of people, fosters extensive ecosystem functions and services, and supports recreational activities such as fishing and swimming. However, warming waters and nutrient-rich runoff that carry human and animal waste create conditions for bacteria such as Enterococci to thrive and reach harmful levels. Testing water for bacteria counts is critical for supporting public health. Often, bacteria testing can be expensive and takes at least 24 hours to produce results, highlighting the importance of finding a more accessible and immediate way for communities to assess water safety the same day. Since 2023, Chesapeake Water Watch (CWW) has investigated Colored Dissolved Organic Matter (CDOM)—decaying plant matter—as a potential proxy for Enterococci counts. Runoff conditions that promote bacteria can also increase CDOM, which can be measured in minutes using a portable fluorometer. As a participatory science project, CWW partners with riverkeepers, local monitoring groups, and volunteers to validate the use of Aquafluors for measuring water quality metrics such as CDOM. Beginning in 2023 with a limited number of bacterial testing sites, expanded in 2025 to multiple tributaries, results indicate that high CDOM levels may indicate high Enterococci counts ($p\text{-value} < .001$). Continued investigation aims to understand CDOM-bacteria relationships to support rapid local water safety assessments.

Veronica Manka'a Tangiri (Prince William Soil and Water Conservation District (PWSWCD), Virginia)

Analysis of the Health and Population of Benthic Macroinvertebrates with Increasing Development: Quantico Creek Watershed, Prince William County, Virginia

Prince William County is Virginia’s second-most populous county and the 10th most diverse county in the United States (Census, 2020), with over 51,000 new residents added between 2015 and 2025. Prince William Forest Park protects a large portion of the Quantico Creek watershed, which scientists use as a benchmark for water quality in the county. Quantico Creek has the highest multi-metric water-quality score in Prince William County and supports rich benthic macroinvertebrate diversity. This work applied data science and machine learning techniques to examine water-quality data from Quantico Creek and assessed how land-use changes associated with development affected the health and diversity of benthic macroinvertebrates amid climate change. The Multi-Metric Score of the Virginia Save Our Streams monitoring protocol, indicating ecological condition, was used with data from 2015–2025 for the South Fork of Quantico Creek. This study highlights the importance of citizen science in stream health studies to support decision-making, underscores the need for meaningful partnerships with the National Parks, local municipalities, and county governments, and emphasizes the integration of machine learning for policies that protect riparian buffers, mitigate water pollution, and manage land-use changes under a changing climate, promoting sustainable development and human health.