Spatiotemporal Modeling of Water Chemistry to Predict Microbiological Events and Agricultural Yield Patterns Across New York State

Alysar Tabet Tatiana Fernandez Ambica Chaki

Professor Huanying (Helen) Gu, Ph.D.
Professor of Computer Science
Associate Dean for Research

New York Institute of Technology

Abstract

Understanding how water quality influences agricultural productivity and microbiological events is critical for sustainable resource management in New York State. This study develops a multi-dataset spatiotemporal deep learning framework that integrates surface-water chemistry, groundwater indicators, microbiological monitoring, and county-level agricultural yield data from 2007–2025. Using monthly aggregated chemical features, Random Forest classifiers were trained to predict microbial events in both drinking-water distribution systems and natural surface waters, as well as to classify interannual yield variability. Results reveal solid predictive relationships between nutrient loading (particularly nitrate and phosphorus) and microbial contamination in rivers and lakes, while turbidity and interannual variability emerge as key drivers of yield class. The proposed pipeline demonstrates how harmonized environmental datasets and spatiotemporal ML can jointly characterize water-quality risks and agricultural outcomes, offering a scalable foundation for statewide water-agriculture system intelligence.

Keywords– Deep Learning, Spatiotemporal Modeling, Water Quality, Agriculture, Agricultural Productivity, New York State

1.Introduction

1.1 Motivation

New York's agricultural economy, valued at over $3.6 billion annually, depends heavily on the quality of its surface and groundwater systems, yet many of the state's lakes and reservoirs remain classified as impaired due to nutrient and contaminant runoff from agricultural activity. The intersection of urban and rural water use highlights a complex feedback loop: agricultural practices influence downstream urban water quality, while urban consumption and wastewater management affect the sustainability of upstream farming. Prior work emphasizes that watershed health is a shared community responsibility, and yet the spatial and temporal dynamics of this relationship remain poorly quantified, limiting our ability to translate water-quality interventions into measurable agricultural and ecological outcomes.

In this study, we develop a multi-dataset, spatiotemporal machine-learning framework to bridge these gaps for New York State. We integrate (i) NYC Department of Environmental Protection distribution-system monitoring for coliform and *E. coli*; (ii) statewide surface-water chemistry and microbiological observations from national water-quality portals; (iii) watershed and aquifer-scale groundwater chemistry from the NYC watershed and Long Island Pine Barrens USGS data releases; and (iv) county-level corn grain yields from USDA NASS. After aggregating water-quality measurements to monthly and annual scales and engineering temperature, turbidity, pH, conductivity, nitrate, phosphorus, and microbial event-rate features, we train Random Forest classifiers to:

(Framework 1) predict microbiological events in drinking-water distribution systems and natural surface waters, and (Framework 2) classify county-year yields into low, medium, and high tertiles.

The modeling results show that nutrient-related variables, particularly nitrate and phosphorus, carry strong signal for predicting surface-water microbial events, while turbidity and interannual variability in water-quality indicators are consistently salient for separating yield classes. Across tasks, machine-learning models substantially outperform simple majority-class baselines, demonstrating that harmonized environmental datasets contain meaningful, exploitable structure linking water quality, microbiology, and agricultural outcomes. By packaging the entire workflow: from data acquisition and feature engineering to cross-validated performance, feature importance, and SHAP-based attribution, into a reproducible pipeline, this work offers a scalable template for statewide water-agriculture intelligence in New York and other data-rich regions.

1.2 Understanding Water Quality Parameters

Water quality is determined by three main factors: biological, physical and chemical. Each plays an essential role in assessing whether water is safe for human consumption, supports ecosystems, and meets industrial standards. Biological factors refer to living organisms found in water such as bacteria, viruses, algae, and protozoa. These microorganisms influence whether water is biologically safe. For instance, when pathogens are present, the water is deemed unsafe to drink, as they can cause illness and disrupt aquatic ecosystems (Agriculture Institute). Physical factors include measurable traits such as temperature, turbidity (cloudiness), color, taste, and odor. These influence not only the usability of water but also its perception – smelly or cloudy water often indicates contamination or imbalance. For example, unpleasant odors may signal the presence of hydrogen sulfide, which, while naturally occurring, can affect soil and water chemistry at high concentrations. Chemical factors describe the water's composition and include properties like pH, dissolved oxygen, nutrients, and heavy metals. A balanced chemical profile is crucial: extreme pH levels can harm aquatic life and interfere with water treatment, while excess metals such as lead or mercury can make water toxic to both humans and animals (Agriculture Institute).

Chemistry serves as the foundation for understanding and interpreting these factors. It enables analysis of elements such as pH, dissolved oxygen, nitrates, phosphates, and heavy metals – all of which are present in our datasets. These measurements help determine whether a body of water meets safety standards set by agencies like the New York State Department of Environmental Conservation (NYSDEC). Even small imbalances or contaminant levels can have far-reaching effects on agricultural productivity and sustainability. Many of these chemical indicators are first noticeable through changes in color, taste, or odor, which often signal underlying issues such as elevated metals or organic matter. For example, a metallic taste or a greenish tint in water can indicate the presence of copper, zinc, or lead, all of which are metals that, in excess, can harm both plants and soil microorganisms essential for nutrient cycling ("Get Informed | Color, Taste, and Odor"). Similarly, odors resembling rotten eggs may point to hydrogen sulfide, which, though naturally occurring, can affect water usability and soil chemistry if concentrations are high. Chemical data can thus reveal contamination trends and help identify areas that may be unsafe for drinking, recreation, or agriculture. In agricultural and coastal regions, water chemistry directly impacts productivity and sustainability. For example, the *2022–2023 Harbor Survey Report* highlights how variations in nutrient concentrations, dissolved oxygen, and contaminants reveal spatial pollution patterns in New York City's waterways (NYC Department of Environmental Protection). Such imbalances, particularly nutrient overloads or runoff, can affect nearby agricultural zones by altering soil salinity and pH, potentially leading to reduced crop yields and ecosystem stress.

Moreover, heavy metal contamination poses one of the most persistent threats to water and soil quality (Zhang et al.). Metals like arsenic, cadmium, and lead accumulate in both soil and plants, disrupting enzymatic processes and entering the food chain . Once bound to soil particles, these pollutants can persist for decades, reducing long-term agricultural productivity and threatening food security. Sustainable agricultural practices in New York, therefore, depend heavily on monitoring and managing these chemical pollutants to ensure that irrigation water remains safe and balanced.

1.3 Scientific Problem

This project asks four core scientific questions: How do chemical conditions within drinking-water distribution systems influence the occurrence of microbiological events?

NYC's distribution network provides high-frequency measurements of residual chlorine, turbidity, fluoride, pH, and temperature across hundreds of sites. Yet it is not well understood which of these parameters best predict short-term coliform and *E. coli* detections, particularly under low-prevalence conditions (≈3.5%). This poses a fundamental challenge: whether routine treatment and distribution chemistry can be used to forecast microbial events at a monthly resolution, and which parameters contribute most strongly to predictive signal.

Which surface-water characteristics are most strongly associated with pathogen risk across rivers, lakes, and harbor environments statewide?

Thousands of physical and chemical measurements from WQP and NYC Harbor monitoring programs capture variability in turbidity, nutrients, conductivity, pH, and temperature. Microbial data (fecal coliform, enterococcus) are sparse, heterogeneous, and collected across multiple sampling programs. This study discovers whether harmonized monthly water-quality indicators predict microbial exceedances at the statewide scale, if nutrient-related variables (particularly nitrate and phosphorus) emerge as universal predictors of microbial events, and how model performance differs between inland waters and coastal/urban systems.

Do water-quality anomalies contain predictive information about interannual agricultural yield variability?

Crop yields are typically modeled using weather, soils, and remote sensing, not water chemistry. This study explores a novel cross-domain question, in whether aggregated surface-water indicators (turbidity, nutrients, temperature, pH), combined with year and county identity, predict whether annual yields fall into low, medium, or high classes. Furthermore, it questions whether adding groundwater chemistry, such as nutrients, PFAS, pesticides, and pharmaceuticals measured in the Long Island Pine Barrens monitoring wells, improve or degrade predictive accuracy.

Which chemical signals are most informative for distinguishing yield classes?

Despite the density of environmental monitoring across New York State, the scientific relationships linking water quality, microbiological contamination, and agricultural performance remain poorly quantified. Existing datasets such as drinking-water distribution chemistry, surface-water nutrient and microbial measurements, groundwater chemistry, and county-level crop yields, are collected by different agencies for different purposes. As a result, New York lacks an integrated, predictive understanding of how chemical and hydrological conditions propagate through the water system to influence public-health risks and agricultural outcomes.

1.4 Limitations of Current Practice

There exists a large volume of environmental and agricultural monitoring data collected in New York State, and yet current analytical practices remain fragmented and insufficient for understanding cross-system interactions. Drinking-water monitoring, surface-water assessments, groundwater chemistry surveys, and agricultural yield statistics are typically analyzed independently, each with its own sampling design, temporal resolution, and quality-control protocols. This siloed structure limits the ability to identify relationships spanning water sources, watersheds, and agricultural production zones.

Existing water-quality models also tend to focus on single parameters or specific regulatory endpoints (e.g., compliance with coliform or enterococcus standards), often using linear statistical frameworks that have difficulty capturing the non-linear thresholds, interactions, and spatial heterogeneity inherent in environmental systems. While machine-learning models have been applied in isolated cases, such as beach "nowcast" systems or local nutrient–HAB models these implementations are rarely statewide, multi-source, or integrated with groundwater data and agricultural outcomes. Similarly, agricultural yield prediction models commonly rely on climate, soil, and remote-sensing indicators but typically exclude upstream water-quality signals, despite evidence that hydrological conditions influence crop stress, nutrient availability, and disease pressures.

Another major limitation is the inconsistency in data harmonization. Water-quality datasets vary widely in units, detection limits, sampling frequencies, and characteristic names. Without standardized processing pipelines, large-scale multi-source analyses remain labor-intensive and prone to errors. Moreover, few existing tools provide explainability frameworks needed to interpret predictions, quantify feature importance, and communicate data-driven insights to water managers and agricultural planners. Finally, the lack of integrated spatiotemporal modeling prevents agencies from leveraging statewide datasets to proactively identify emerging risks or support cross-sector decision-making.

1.5 Contributions of this Work

This study addresses these gaps by developing a unified, multi-dataset spatiotemporal machine-learning framework that links water chemistry, microbial contamination, and agricultural outcomes across New York State. The contributions are both methodological and scientific:

- A harmonized environmental data pipeline for New York State

We consolidate and standardize drinking-water distribution chemistry (NYC DEP), statewide surface-water quality and microbiology (WQP and NYC Harbor), groundwater chemistry from USGS aquifer studies, and USDA NASS county-level crop yields. This pipeline resolves unit inconsistencies, non-detect handling, temporal aggregation, sampling heterogeneity, and site metadata integration, enabling cross-domain analyses at monthly and annual scales.

- Predictive modeling of microbial events using routine chemical indicators

For both drinking-water distribution systems and natural surface waters, we train Random Forest models that significantly outperform majority-class baselines. In NYC's distribution network (Framework 1, Model 1), residual chlorine, turbidity, and temperature emerge as key predictors of coliform and *E. coli* events. In statewide surface waters (Framework 2, Model 1), nutrient indicators - especially nitrate and phosphorus - exhibit the strongest predictive signal for fecal contamination. These models provide actionable insights for monitoring optimization and risk-based sampling strategies.

- A novel linkage between water-quality patterns and county-level crop yields

By combining water-quality features, interannual variability, and groundwater chemistry from the Pine Barrens aquifer system, we demonstrate that water-related indicators carry predictive information about agricultural yield classes (Framework 2, Models 1 and 2). Turbidity, nutrient signals, and groundwater-derived variables contribute disproportionately to classification performance, revealing potential hydro-environmental pathways influencing agricultural outcomes.

- Explainability and feature attribution for environmental decision-making

We integrate SHAP-based feature attribution to identify the drivers behind microbial-event and yield predictions. This explainability component allows stakeholders to understand why models succeed, assess variable importance (e.g., nitrate dominance in surface-water microbial modeling), and interpret groundwater contaminant contributions in yield classification.

- A scalable blueprint for statewide water–agriculture intelligence

Finally, the work offers a generalizable framework: combining automated data acquisition, spatiotemporal aggregation, machine learning, and interpretability, that can be applied to other states or expanded to include remote sensing, climate projections, or mechanistic hydrological models. The approach opens the door to integrated early-warning systems for microbial risk, nutrient management, and agricultural planning.

2.Related Work

2.1 Water Quality and Microbial Contamination Modeling

Work on modeling microbial contamination in surface and drinking waters has grown rapidly, especially around fecal-indicator bacteria, cyanobacteria, and nutrient-driven blooms. Early “nowcast” models for recreational waters used regression on rainfall, turbidity, and antecedent conditions to forecast E. coli and enterococci at beaches and river sites. More recent work layers in machine-learning models (Random Forests, gradient boosting, deep learning) and remote sensing to improve spatiotemporal prediction of microbial risk and harmful algal blooms.

Within and near New York, there is a strong focus on Lake Champlain (Vermont Department of Health), the Hudson/estuarine system, and Long Island groundwater. Studies in Lake Champlain use multi-year buoy, in-situ, and meteorological data to train machine-learning classifiers and regression models for cyanobacterial bloom onset and intensity, explicitly linking nutrient enrichment, temperature, and stratification to bloom dynamics.

Estuarine work along the Hudson and New York Harbor has examined enterococci/E. coli dynamics along salinity and urbanization gradients, often using statistical or semi-mechanistic models to relate microbial indicators to tides, storm events, and wastewater inputs. Parallel efforts in (Marrone, Banerjee and Talapatra) Long Island have focused on nitrate-contaminated groundwater and its implications for surface-water eutrophication, providing context for the Pine Barrens datasets used (Department of Environmental Conservation).

2.2 Agricultural Yield Prediction

Crop-yield prediction has a long history in agronomy, starting with regression models that link yields to weather indices and gradually expanding to incorporate remote sensing, soil properties, and management data. Over the last decade, county- and field-scale yield models based on Random Forests, gradient boosting, and deep neural networks have demonstrated clear gains over linear baselines, particularly when fusing multi-source data (Salami, Babamuratov and Ayyappan). This is what led to the decision of using the Random Forest model, due to its precedence.

While many studies are national or global, several use USDA NASS county-level yield data – the same backbone used in this project – and emphasize both cross-sectional and temporal generalization. In the Northeastern US and New York, modeling efforts often focus on corn and forage crops under cold-temperate climate and variable soils, using statistical or process-based models to relate yields to growing-season thermal time, precipitation, and soil moisture; this study extends this by coupling yields to water-quality and microbial-event indicators (Kulyal and Saxena).

2.3 Machine Learning in Environmental Science

Across environmental science, ML has become a core tool for prediction, data fusion, and pattern discovery in complex spatiotemporal systems. Recent reviews document adoption in hydrology, water-quality modeling, ecology, and climate science, with applications ranging from streamflow forecasting and water temperature prediction to species distribution modeling and air-quality nowcasting.

A major theme is “physics-guided” or “theory-guided” machine learning (Karpatne, Atluri and Faghmous), which blends data-driven models with physical constraints or process knowledge. This direction responds to concerns about extrapolation, non-stationarity, and interpretability. Studies on river-network temperature, streamflow, and hydrologic fluxes show that combining graph neural networks or RNNs with hydrologic equations can outperform both standalone ML and purely process-based models. SHAP, permutation importance, and related explainability tools are now standard to interpret feature contributions, similar to how we analyze nitrate, turbidity, and Pine Barrens contaminants in the models.

3. Data and Sources

All raw datasets were obtained from public agencies and transformed into a small number of modeling-ready tables that feed the four Random Forest models described in both the Introduction and Methodology (Frameworks 1 and 2).

3.1 Drinking-Water Distribution System Chemistry and Microbiology (NYC DEP)

The first data stream characterizes treated drinking water as it moves through New York City's distribution system. The New York City Department of Environmental Protection (NYC DEP) routinely monitors physical–chemical and microbiological indicators (e.g., residual chlorine, turbidity, fluoride, temperature, pH, total coliform, *E. coli*) at hundreds of locations across the five boroughs, as documented in its annual Drinking Water Supply and Quality reports.

This project works from the NYC DEP distribution monitoring dataset that underlies these reports (accessed via NYC's open data infrastructure and associated metadata API). The raw table contains site identifiers and location/context (site ID, facility name, borough, sometimes address or zone), sampling date and time. Disinfectant and basic chemistry residual chlorine (free/total), turbidity, fluoride, temperature, pH, and related fields. Microbiological indicators: total coliform and *E. coli* presence/absence or counts, plus QA flags.

From this, a monthly distribution panel is built, with each row corresponding to one distribution site × one year-month. For each site–month the mean residual chlorine, turbidity, fluoride, temperature, pH, as well as a binary label y are computed indicating whether at least one coliform or *E. coli* event occurred in that month (using the DEP microbiological measurements and flags).After quality filtering, the modeling subset used in Framework 1 Model 1 has 3,280 site-month rows, with microbial event prevalence ~3.5% (mostly non-event months, as expected).

This aggregated table is saved as dist_monthly and exported as dist_monthly.json for use in the RIW frontend.

3.2 Surface Waters (USGS WQP, NYC DEP Harbor)

The second data stream captures surface-water quality in rivers, lakes, and coastal waters across New York State. A unified surface-water table constructed from the Water Quality Portal (WQP), a national integrator that aggregates USGS, EPA, and state/tribal monitoring data, including New York results for nutrients, turbidity, temperature, pH, and microbial indicators. NYC DEP Harbor Survey and related coastal monitoring programs, which provide long-term nutrient and bacteria time series for New York Harbor and adjacent estuarine waters (accessed via NYC DEP's open data and reports).

Starting from raw records with station IDs, coordinates, sample dates, analyte names, measurement values, and detection flags, the project filters to New York surface-water sites within the WQP queries (state code US:36) and to NYC Harbor stations from DEP. Then it standardize analytes and units into canonical variables, using mapping dictionaries (such as Temperature → temp_C).. It handles non-detects and special codes by parsing WQP codes such as (D), (L), (Z), (NA) and treating them as missing or low-concentration surrogates where appropriate.

It constructs monthly aggregates. For each site × month, computes mean temp_C, turbidity_ntu, ph, cond_uscm, nitrate_mgL, phosphorus_mgL, builds a surface microbial label y_surface representing the monthly rate of microbial "events" (e.g., fraction of samples exceeding coliform/*E. coli*/enterococcus thresholds) at that site and month.

To align with the agricultural record, the months are strictly from 1970 onward and export a reduced version surface_monthly (saved as surface_monthly.json) for visualization.

3.3Groundwater and Contaminant Datasets (USGS ScienceBase)

The third data stream contributes groundwater and contaminant context, especially from Long Island's vulnerable sand and gravel aquifer systems. The Long Island Pine Barrens are a large pitch-pine and scrub-oak forest and recharge area located primarily in central and eastern Suffolk County on Long Island, New York. The region overlies critical sole-source aquifers and is designated as a preserved protection area due to its importance for regional drinking-water supply and ecological value.

USGS has released a series of groundwater-quality data releases for the Long Island Pine Barrens region (e.g., 2020–2022 ScienceBase Excel releases) that include nutrients, major ions, trace elements, pesticides, pharmaceuticals, and groundwater-level measurements at monitoring wells across this aquifer system.

The project ingests multiple USGS/ScienceBase datasets:

- USGS Long Island Pine Barrens data releases (2020, 2021, 2022): each year's Excel workbook includes sheets such as:
    - *Field* (water temperature, pH, specific conductance, dissolved oxygen, etc.),
    - *Nutrients* (nitrate, nitrite, ammonia, phosphorus species),
    - *Inorganics* (major ions and trace metals),
    - *Pesticides*,
- An NYC DEP Watershed Water Quality dataset, providing chemistry for upstate reservoir and watershed sites feeding the NYC water supply (used in Model 2 as an additional annual environmental context panel).

Because these datasets have metadata and units baked into the first few Excel rows, the notebook skips metadata and unit rows and then parses the data tables. A helper function load_pine_sheet handles this consistently across years and sheets.

3.4 Agricultural Yield Data (USDA NASS)

Agricultural yield outcomes come from the USDA National Agricultural Statistics Service (NASS), which provides county-level production, acreage, and yield statistics via its Quick Stats system.

a CSV export of the USDA NASS regional crop yield data that contains:

- State, county, and geographic level (STATE, COUNTY, GEO LEVEL)
- Year (YEARCommodity (COMMODITY, e.g., "CORN")
- Data item (e.g., "CORN, GRAIN - YIELD, MEASURED IN BU / ACRE")
- Reported value (VALUE as a string with commas and code values),
- Numeric codes and descriptors for units and statistics.

The USDA NASS dataset was first filtered to retain only records corresponding to New York State at the county level. This was accomplished by selecting rows where STATE = "NEW YORK" and GEO LEVEL = "COUNTY", ensuring that all subsequent analyses reflected the spatial granularity relevant to county-level yield prediction. Once the geographic scope was defined, the numeric yield values required extensive cleaning. The raw VALUE field contains commas and occasionally non-numeric codes such as "(D)", "(L)", "(Z)", "(NA)", or empty strings, which are used by NASS to denote suppressed or unavailable data. All commas were removed and the remaining strings were converted into a numeric field, VALUE_NUM. Any rows carrying non-numeric codes were treated as missing observations and discarded, producing a fully numeric yield dataset suitable for modeling.

Commodity selection and yield extraction proceeded by narrowing the dataset to commodities with consistent multiyear, multi-county reporting. Although several commodities were explored including barley, beans, hay, hemp, and herbs, the final analysis focused on corn grain yield because it is widely cultivated in New York and exhibits reliable annual reporting across counties. The dataset was therefore restricted to entries where COMMODITY = "CORN" and where the DATA ITEM explicitly referred to yield. Because NASS distinguishes grain yield, silage yield, and other yield types, the DATA ITEM field was simplified to classify yield entries as grain, silage, or other. When multiple yield types were available for the same county and year, grain yield was preferentially selected; otherwise, all available yield entries were retained. For each county-year pair, the mean of VALUE_NUM was taken to produce a single continuous yield estimate. County names were then normalized by removing inconsistent formatting; including trailing " COUNTY" strings, and converting them to a standardized uppercase form, resulting in a clean COUNTY_NAME field.

To enable classification-based modeling, continuous corn grain yield values were transformed into categorical yield classes. The 33rd and 66th percentiles of all county-year yields were computed to establish thresholds for low, medium, and high performance. Observations falling below the lower threshold were labeled as low yield, those between the lower and upper thresholds were labeled as medium yield, and those exceeding the upper threshold were labeled as high yield. This discrete label, y_yield, was assigned to each county-year observation and forms the primary dependent variable used in Framework 2. The resulting dataset: comprising YEAR, COUNTY_NAME, COMMODITY, continuous YIELD, and the categorical label y_yield, constitutes the final yield panel and is also exported as yield_panel.json for use in the RIW frontend.

3.5 Pre-Processing and Feature Engineering

Across all datasets, pre-processing and feature engineering followed consistent principles designed to harmonize heterogeneous environmental records into a unified analytical structure. All date fields, whether representing individual sampling events or pre-aggregated month-start indicators, were parsed as true datetime objects. For the microbiology and water-quality models in Framework 1, both the NYC drinking-water distribution records and statewide surface-water records were aggregated to the monthly scale at the monitoring-site level. Monthly keys were defined using either a year_month field for distribution sites or a month_start field for surface-water stations. In contrast, the yield models in Framework 2 required annual-scale features; thus, surface-water and Pine Barrens groundwater datasets were aggregated by year. Surface-water features were additionally restricted to the growing season, defined as April through September, ensuring that only agriculturally relevant water-quality conditions informed the annual predictors. Pine Barrens groundwater datasets were already provided on an annual basis and therefore required no temporal downscaling.

Chemical measurements across datasets used diverse naming conventions, units, and quality-assurance flags, necessitating extensive standardization. All physicochemical variables were mapped into a consistent schema, yielding canonical feature names such as temp_C_mean, turbidity_ntu_mean, ph_mean, cond_uscm_mean, nitrate_mgL_mean, and phosphorus_mgL_mean. Non-detects and QA annotations, particularly those originating from WQP (such as "(D)", "(L)", "(Z)", and "(NA)"), were parsed and converted into missing values to prevent misinterpretation of censored observations. Where datasets differed in measurement units, values were converted into shared units using the metadata provided by WQP and USGS sources: for example, normalizing conductivity to microsiemens per centimeter, nitrate to milligrams per liter, and temperature to degrees Celsius.

Outcome labels were constructed in parallel across the various modeling frameworks. In Framework 1 Model 1, a binary label y was created to indicate whether a monitored distribution site experienced at least one coliform or *E. coli* detection within a given month. For Framework 1 Model 2, surface-water microbial outcomes were expressed as a monthly microbial event rate, defined as the proportion of samples in a month exceeding regulatory or threshold levels for bacterial indicators. This continuous or binary microbial metric, y_surface, served as both a target variable for microbial-event modeling and a precursor to the annual aggregate used in Framework 2. Yield labels in Framework 2 were handled as described above, resulting in the categorical y_yield variable.

Feature engineering proceeded by constructing model-specific predictor sets aligned with the hypotheses underlying each framework. For the NYC distribution microbiology model in Framework 1 Model 1, the feature vector consisted of monthly averages of residual chlorine, turbidity, fluoride, temperature, and pH. The surface-water microbiology model in Framework 1 Model 2 used monthly mean turbidity, temperature, pH, conductivity, nitrate, and phosphorus. For the first yield model in Framework 2, surface-water features were aggregated to annual values. Specifically, growing-season means of temperature, turbidity, pH, conductivity, nitrate, and phosphorus, along with an annualized microbial-event rate termed surface_event_rate were aggregated. These were joined to each county-year yield observation to produce the modeling table. The second yield model in Framework 2 integrated the extensive Pine Barrens groundwater and contaminant dataset; all features beginning with the pine prefix including field parameters, nutrients, inorganics, pesticides, pharmaceuticals, and groundwater levelswere merged into the yield panel to create a high-dimensional annual environmental table. All Pine-derived variables served as predictors for yield classification.

3.6 Summary of Final Modeling Tables

By the end of pre-processing, the project's modeling pipeline was organized around a set of carefully structured tables, each representing a different spatial and temporal scale. The first of these, dist_monthly, contains NYC distribution monitoring records aggregated by site and month. It includes monthly mean values for residual chlorine, turbidity, fluoride, temperature, and pH, along with a binary indicator denoting whether a microbial event occurred during that month. This table forms the basis for Framework 1 Model 1, which predicts microbial events in the distribution system.

A second major table, surface_monthly (and its cleaned and unified form, surface_unified_monthly), aggregates statewide surface-water data by site and month. It includes growing-season measurements of temperature, turbidity, pH, conductivity, nitrate, and phosphorus, as well as a microbial event rate computed from fecal-indicator bacteria observations. This dataset serves both as the direct input for Framework 1 Model 2 and as the raw material for annual water-quality features in Framework 2.

The central agricultural dataset is the panel table, which contains one record per county and year. Each record includes the county name, year, selected commodity (corn), continuous yield, and the categorical yield label derived from percentile thresholds. This table supports both yield models in Framework 2 and is also exported in JSON form for visualization.

To evaluate how surface-water conditions influence agricultural outcomes, the model_df table was constructed by joining the annual surface-water aggregates to each county-year yield record. This table includes the year; annual water-quality means for temperature, turbidity, pH, conductivity, nitrate, and phosphorus; the annual microbial-event rate; and the corresponding yield class. It is used in Framework 2 Model 1.

Finally, the most comprehensive table, yield_env, combines the yield panel with the Pine Barrens environmental dataset. This table contains all Pine-derived annual variables, field measurements, nutrients, inorganics, pesticides, pharmaceuticals, and groundwater levels, alongside the yield class. It provides the high-dimensional feature space used in Framework 2 Model 2, enabling an assessment of whether subsurface contaminant and aquifer conditions contribute measurable predictive power for agricultural yield classification.

4.Methodology

The methodological design of this study centers on constructing a unified spatiotemporal framework that links water-quality dynamics, microbiological events, and agricultural yields across New York State. Heterogeneous datasets from drinking-water distribution networks, surface waters, groundwater systems, and county-level crop statistics are harmonized into consistent monthly and annual representations. These representations are then used to train a sequence of supervised learning models that predict microbiological events and classify yield variability. Random Forest classifiers, combined with stratified cross-validation and baseline dummy models, provide a robust and interpretable modeling backbone. Model interpretation relies on both impurity-based feature importance and SHAP-based attribution, and all key outputs are exported as tidy tables and graph-ready JSON to support interactive mapping and visualization.

4.1 Spatiotemporal Aggregation Framework

The core of the methodology is a spatiotemporal aggregation framework that reconciles datasets collected on different schedules, at different spatial resolutions, and with varying quality control rules. For the microbiological models in Framework 1, the unit of analysis is the monitoring site at monthly resolution. NYC drinking-water distribution measurements from NYC DEP are aggregated by distribution site and year_month, while surface-water measurements from USGS WQP and the NYC DEP Harbor Survey are aggregated by surface-water station and month_start. Within each monthly cell, all available physicochemical measurements are averaged, and microbiological observations are summarized as binary events or event rates depending on the model. This results in two tidy tables: one for distribution-site months and one for surface-site months, each with a consistent set of predictor variables and labels.

For the agricultural yield models in Framework 2, the temporal scale shifts to annual resolution, reflecting that crop yields are reported yearly at the county level. Surface-water features, originally aggregated at the monthly site level, are further summarized into annual, statewide means restricted to the growing season (April–September). This step ensures that only agriculturally relevant hydrological conditions inform the yield models. In parallel, the Long Island Pine Barrens groundwater datasets from USGS ScienceBase are processed into annual panels, with chemical, pharmaceutical, pesticide, and groundwater-level variables aggregated by year. These annual water-quality and environmental tables are joined to the county-year yield panel, producing integrated modeling tables where each row represents a New York county in a given year, enriched with interannual surface-water or Pine Barrens environmental signals.

Throughout the framework, temporal alignment is enforced using explicit YEAR, year_month, and month_start fields, and spatial alignment is achieved via standardization of county names and consistent treatment of site identifiers. This design enables cross-scale analyses in which fine-grained microbial patterns are summarized into annual environmental indicators that can be related to regional agricultural outcomes.

4.2 Microbiological Event Detection Models

Microbiological event detection is implemented using supervised binary classification models trained on monthly site-level records. For both the distribution and surface-water settings, Random Forest classifiers are paired with stratified k-fold cross-validation (with five folds, shuffle enabled, and a fixed random seed) to provide stable performance estimates and to preserve the observed class distribution in each fold. A dummy baseline classifier that always predicts the majority class serves as a benchmark to quantify the added value of environmental predictors.

Class imbalance, especially in the distribution system where microbial events are rare, is addressed through the use of class-weighting (class_weight="balanced" or "balanced_subsample") within the Random Forest models. Model performance is evaluated using the area under the receiver operating characteristic curve (ROC AUC) and the area under the precision–recall curve (PR AUC). ROC AUC characterizes the model’s ability to rank event versus non-event months across thresholds, while PR AUC is particularly informative under strong class imbalance, as it focuses on performance in the minority (event) class. For each model, mean and standard deviation of ROC AUC and PR AUC across folds are reported to capture both central performance and variability.

4.2.1 Distribution System

In the NYC drinking-water distribution system, the modeling target is a binary monthly indicator of microbiological events, defined as at least one detection of coliform or *Escherichia coli* at a given distribution site in a given month. The feature vector for each site–month consists of monthly mean residual chlorine, turbidity, fluoride, water temperature, and pH, all derived from operational and regulatory monitoring data. After filtering out rows with missing labels or missing values in all predictors, the final dist_monthly table contains 3,280 site–month observations, with a positive-event prevalence of approximately 3.5%.

The Random Forest classifier for this setting uses 300 trees, no fixed maximum depth, and a minimum leaf size of one, with class weights adjusted to compensate for the rarity of microbial events. In five-fold stratified cross-validation, the model achieves a mean ROC AUC of approximately 0.736 (± 0.017) and a mean PR AUC of about 0.080 (± 0.014). The dummy baseline, which effectively predicts "no event" for all months, remains at a ROC AUC of 0.5 (random ranking) and a PR AUC close to the event prevalence (~0.035). These results indicate that monthly distribution-system chemistry contains a detectable, non-trivial signal for predicting microbial events and that the Random Forest substantially outperforms a prevalence-based baseline.

4.2.2 Surface Waters

For surface waters, the microbiological modeling focuses on predicting a monthly microbial event rate at each surface site, defined as the fraction of samples in that month exceeding a specified threshold for fecal-indicator bacteria. The feature vector for each site–month includes canonical physicochemical variables: temperature, turbidity, pH, specific conductance, nitrate, and phosphorus, all harmonized and expressed as monthly means. As in the distribution model, records with missing labels or fully missing predictor vectors are excluded to ensure numerical stability.

The Random Forest model for surface water uses a similar configuration with 300 trees and class weighting, but applied to a larger and more heterogeneous statewide dataset that includes both inland WQP sites and NYC Harbor stations. Model performance, evaluated via five-fold stratified cross-validation, shows strong discriminative capacity, with a mean ROC AUC of approximately 0.897 and a mean PR AUC of roughly 0.674. These values indicate a strong predictive relationship between nutrient enrichment (particularly nitrate and phosphorus), turbidity, and the occurrence of microbiological events in New York rivers, lakes, and estuarine waters. Feature importance analysis confirms the dominant role of nitrate, followed by phosphorus and turbidity, aligning with the conceptual understanding that nutrient loading and suspended particulate matter promote microbial proliferation and transport.

4.3 Agricultural Yield Modeling

The agricultural component of the methodology comprises two complementary yield-classification models, both built on the county-year corn yield panel but incorporating different environmental covariates. In both cases, the target variable is a three-class label representing low, medium, and high corn grain yield based on tertiles of the continuous yield distribution across all counties and years. Random Forest classifiers are again used as the primary modeling algorithm, with stratified five-fold cross-validation and a dummy majority-class baseline; performance is evaluated using macro-averaged F1 score and macro-averaged ROC AUC to give equal weight to each yield class.

The first yield model (Framework 2, Model 1) integrates annual surface-water context. Annual growing-season means of surface-water temperature, turbidity, pH, specific conductance, nitrate, and phosphorus, along with an annual microbial-event rate (surface_event_rate), are joined to the yield panel on YEAR. The resulting model_df table consists of county-year records that carry both yield labels and statewide water-quality indicators. The Random Forest classifier, with 300 estimators and class weighting, attains a macro F1 score of approximately 0.512 (± 0.033) and a macro ROC AUC of about 0.689 (± 0.021), substantially above the dummy baseline (macro F1 ~0.169, macro AUC 0.5). These results suggest that interannual variability in surface-water conditions and microbial stress is meaningfully associated with differences in yield class across New York counties.

The second yield model (Framework 2, Model 2) extends this approach by incorporating high-dimensional groundwater and contaminant features derived from the Long Island Pine Barrens region. Annual panels from the Pine Barrens data release, including field measurements, nutrient concentrations, inorganics, pesticides, pharmaceuticals, and groundwater levels, are merged into the yield panel to form the yield_env table. All columns with names beginning with the pine prefix constitute the predictor set, capturing a rich representation of aquifer and contaminant conditions over time. After removing records where all Pine features are missing, a Random Forest classifier with 300 trees and class weighting is trained and evaluated using stratified five-fold cross-validation. The model achieves a macro F1 score of approximately 0.395 (± 0.109) and a macro ROC AUC of about 0.669 (± 0.059), outperforming the dummy baseline (macro F1 ~0.194, macro AUC 0.5) despite the high dimensionality and relatively limited temporal span (primarily 2020–2022). This configuration tests whether detailed subsurface and contaminant signals from an important recharge region add discriminative power for yield classification beyond what can be inferred from surface-water conditions alone.

4.4 Model Interpretation

Model interpretability is addressed through a combination of Random Forest feature importance metrics and SHAP-based feature attribution, with the goal of identifying which environmental variables most strongly influence predictions of microbiological events and yield classes. For each Random Forest model, impurity-based feature importance values are extracted and visualized as horizontal bar charts, ranking predictors by their contribution to reducing classification error across the ensemble of trees. In the surface-water microbiology model, these plots highlight nitrate and phosphorus as leading drivers of microbial event risk, followed by turbidity and conductivity, reflecting the strong link between nutrient loading, particulate matter, and bacterial proliferation. In the yield models, annual turbidity and microbial-event rate emerge as influential surface-water predictors, suggesting that interannual patterns of water quality and microbiological stress are informative for crop performance.

For the Pine Barrens–enriched yield model, SHAP (SHapley Additive exPlanations) is used to obtain a more granular, class-specific interpretation. A TreeExplainer is fit on the trained Random Forest, and SHAP values are computed for a representative subset of the feature space. For multi-class classification, SHAP values are examined separately for the high-yield class, isolating the patterns most associated with successful yield outcomes. Summary plots in bar form illustrate the top Pine-derived features ranked by their mean absolute SHAP value, indicating which groundwater, nutrient, inorganic, pesticide, pharmaceutical, and groundwater-level variables exert the largest marginal influence on predicting high-yield versus lower-yield years. Shortened feature labels are constructed to improve readability, while the corresponding full variable names are preserved in exported tables to maintain traceability back to the original USGS data schema.

In addition to aggregate importance metrics, the SHAP export includes a ranked table of the top 20 Pine features for the high-yield class, saved as both CSV and PNG. This table is intended to support detailed interpretation in the results and discussion sections, enabling examination of whether specific contaminants, nutrient patterns, or groundwater-level dynamics consistently align with higher or lower yield classes. Together, these interpretability tools provide evidence that the models are learning physically and agronomically plausible relationships rather than relying solely on spurious correlations.

4.5 Mapping and Visualization

To support interactive exploration and communication of the modeling results, the pipeline exports key modeling tables and graph series to JSON and image formats suitable for a web-based visualization interface. For Framework 1, the dist_monthly and surface_monthly tables are written to dist_monthly.json and surface_monthly.json, respectively. These exports preserve site identifiers, dates, monthly mean physicochemical variables, and microbial labels or event rates, enabling the frontend to render time series plots, event timelines, and map overlays that highlight spatial patterns in microbial risk across both the NYC distribution network and statewide surface waters.

For Framework 2, the county-year corn yield panel is exported as yield_panel.json, including year, county name, continuous yield, and the discrete y_yield class. This file feeds into map visualizations where New York counties can be symbolized by yield class or by interannual yield anomalies. Additional JSON files capture graph-ready series derived from the modeling code, such as annual mean corn yield over time, annual surface microbial event rates, and county-year scatter data linking yield to microbial stress indicators. These series are used to generate line charts and scatter plots that contextualize the machine-learning outputs with direct empirical trends.

Collectively, the mapping and visualization layer allows the integrated water–agriculture system to be explored in an interactive, spatially explicit way. Users can view how microbial event hotspots in rivers, lakes, and distribution systems correspond to regions of persistent low or high yield; track changes in nutrient and microbial conditions over time; and interrogate model predictions alongside the underlying environmental and agricultural data. This tight linkage between modeling, interpretation, and visualization is intended to make the analytical outputs more transparent, actionable, and accessible to researchers, practitioners, and decision-makers concerned with New York's coupled water and food systems.

5.Results

5.1 Drinking Water

Microbiological events in the NYC drinking-water distribution system were rare, occurring in only approximately 4% of all site–month observations. This low prevalence is clearly visible when examining monthly trends in event occurrence.

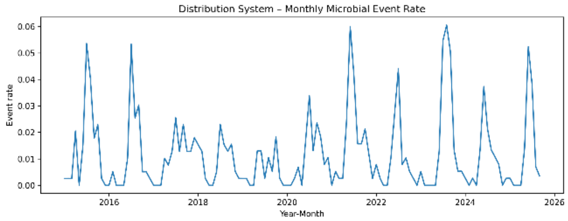


Figure 1 – Monthly Microbial Event Rate

Figure 1 highlights a noisy and predictably low background, consistent with expectations for a well-regulated, disinfected municipal system. Daily-scale variability also reveals sporadic spikes in positive samples, which often correspond to individual site anomalies or isolated operational fluctuations.

Understanding the chemical environment underlying these events requires examining the distributions of residual chlorine, turbidity, pH, fluoride, and temperature across all monitored sites. \ Chlorine residuals exhibit expected seasonal modulation, while turbidity and temperature vary more widely by site and borough. To further illustrate the variability in disinfectant behavior, showing how chlorine trends shift over the year. Similarly, turbidity patterns vary significantly by location, and their relationship to microbial positivity is of particular interest to demonstrate this.

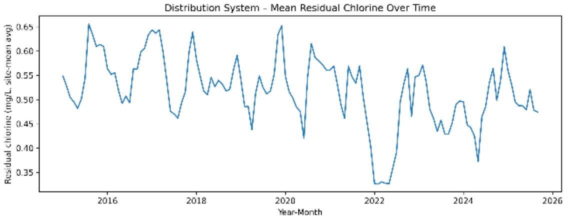
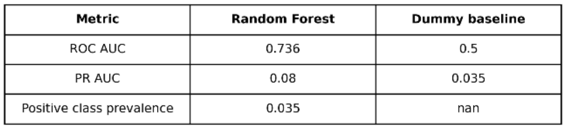


Figure 2- Chlorine Residuals Time Series

Preliminary visual analysis suggests that microbial events cluster under conditions of elevated turbidity and reduced chlorine residual. Scatterplots make this relationship more explicit. When highlighting only event months, the pattern becomes clearer, turbidity levels are often elevated while disinfectant intensity is lower.

The Random Forest classifier trained on monthly chemistry performed significantly better than the dummy baseline, achieving a mean ROC AUC of 0.736 and PR AUC of 0.0798 across five folds. This is notable given the highly imbalanced nature of the dataset and shows that microbial events, although rare, are predictable using monthly aggregated chemistry variables. This finding has practical implications: it suggests that existing regulatory monitoring data can be leveraged for proactive microbial risk profiling, even without specialized sensors, high-frequency measurements, or hydraulic simulations. While the model is not intended for operational real-time decision-making, it provides empirical evidence that low-frequency monitoring datasets contain actionable signatures that correspond to microbial vulnerability within large distribution systems.



Metric	Random Forest	Dummy baseline
ROC AUC	0.736	0.5
PR AUC	0.08	0.035
Positive class prevalence	0.035	nan

Figure 3-Training Results for Microbiology Prediction Framework: Coliform and E.Coli in NYC Drinking Water

The model effectively ranked the rare event months above non-event months, even though precise prediction at monthly scale remains inherently difficult in low-event settings

5.2 Surface Waters

Surface-water systems in New York display far greater variability in water-quality parameters than the controlled environment of the distribution system. Across the dataset, nitrate, phosphorus, turbidity, conductivity, temperature, and pH fluctuate significantly across rivers, lakes, and estuarine waters. These gradients reflect both natural watershed heterogeneity and anthropogenic pressures such as agriculture, wastewater inputs, stormwater pulses, and seasonal stratification.

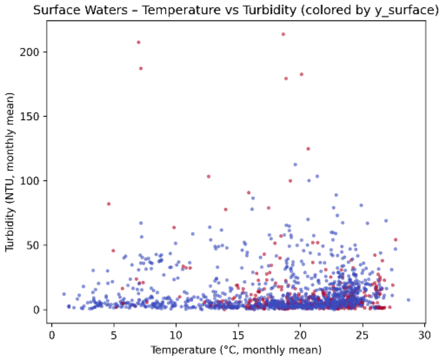


Figure 4 – Microbial Events plotted in Temp vs Turbidity

Microbial event rates, computed monthly as the proportion of samples exceeding bacterial thresholds, show strong temporal structure and distinct differences between inland and coastal waterbodies. Estuarine systems, particularly those influenced by combined sewer overflows or high recreational use, exhibit higher and more variable event rates.

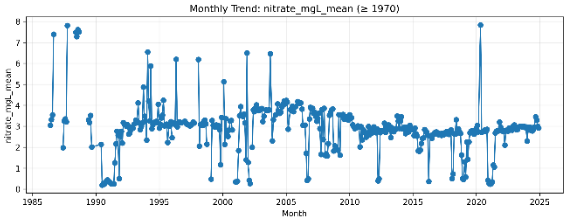


Figure 5 – Nitrate Mean Time Series (from 1970)

Despite this complexity, the Random Forest classifier trained on monthly aggregated surface-water features performed exceptionally well. The model achieved a mean ROC AUC of 0.8966 and a mean PR AUC of 0.6742, far surpassing the distribution-system model. This improvement reflects both a stronger overall signal and a more balanced microbial event distribution across surface waters. The high PR AUC demonstrates that the model excels at identifying months with elevated microbial contamination, a critical capability for public-health protection, recreational advisories, and watershed management.

Feature importance results indicate that nitrate is the strongest predictor of microbial exceedance, followed closely by ph and turbidity.

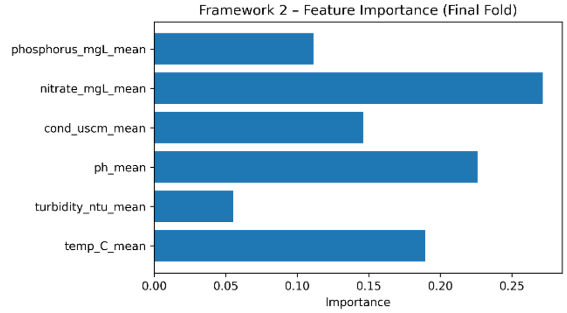
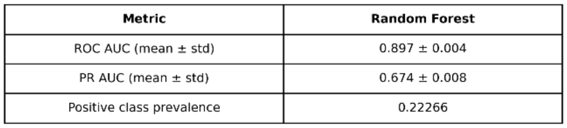


Figure 6 – Feature Importance

This aligns with the conceptual understanding that nutrient enrichment promotes microbial growth directly and indirectly through increased organic matter, while turbidity often reflects runoff, resuspension, and particle-mediated transport of bacteria. Conductivity and temperature also contribute meaningfully, highlighting the influence of watershed hydrochemistry and seasonal thermal regimes. The dominance of these variables is consistent across folds, underscoring the statistical robustness of the relationship between nutrient loading and microbial exceedances.



Metric	Random Forest
ROC AUC (mean ± std)	0.897 ± 0.004
PR AUC (mean ± std)	0.674 ± 0.008
Positive class prevalence	0.22266

Figure 7 – Training Results for Microbiology Prediction Framework: Fecal Coliform and Enterococcus in NY Rivers, Lakes and Harbor Waters

Temporal trend analyses further contextualize model behavior. Nitrate concentrations and turbidity levels show upward trends in several regions, particularly in watersheds influenced by urbanization or intensive agriculture. Microbial event rates appear to co-vary with these trends, indicating system-wide sensitivity to environmental change. Together, these findings demonstrate that New York's surface waters exhibit strong, learnable relationships between physicochemical conditions and microbial exceedances, and that routine monitoring datasets already capture key signals necessary for predictive modeling on monthly timescales.

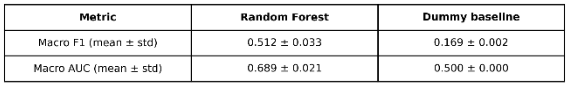
5.3 Agricultural Yield Model

5.3.1 Surface Water

Corn grain yields across New York counties show substantial interannual variability, a product of fluctuating climate conditions, management strategies, soil properties, and hydrological stressors. When continuous yield values were converted into low-, medium-, and high-yield classes based on percentile thresholds, the resulting distribution created a balanced and interpretable target for multi-class classification. Annual surface-water characteristics, aggregated across the state, exhibit their own interannual variability, with fluctuations in temperature, turbidity, nitrate, phosphorus, and microbial event rates reflecting broader environmental conditions.

The Random Forest classifier integrating these annual hydrological and chemical indicators achieved a macro-averaged F1 score of 0.512 (± 0.033) and a macro ROC AUC of 0.689 (± 0.021). These metrics far exceed those of the dummy classifier, demonstrating that surface-water conditions contain meaningful information about agricultural yield outcomes. Although water-quality variables are not direct determinants of crop yield, they serve as proxies for hydrological regime, climatic anomalies, runoff intensity, and watershed-level environmental stress, all of which can influence agricultural performance. The moderate but consistent predictive signal indicates that annual water chemistry aligns with broader environmental patterns that affect crop growth at county scale.

Feature importance analysis shows that turbidity and microbial event rate are among the strongest predictors, followed by nitrate and conductivity. These patterns suggest that years with elevated watershed disturbance, nutrient loading, or microbial stress broadly coincide with yield-limiting conditions such as excess rainfall, soil saturation, or runoff-driven nutrient imbalance. While the relationships are indirect, the alignment between hydrological variability and crop outcomes demonstrates a potentially under-explored linkage between agricultural systems and watershed-scale water quality.



Metric	Random Forest	Dummy baseline
Macro F1 (mean ± std)	0.512 ± 0.033	0.169 ± 0.002
Macro AUC (mean ± std)	0.689 ± 0.021	0.500 ± 0.000

Figure 8 – Training Results for Crop Yield Prediction Framework: Surface-Water Quality and Microbial Event Rates

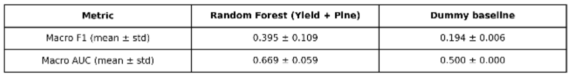
This model demonstrates that statewide water-quality datasets, when aggregated and simplified, can meaningfully augment agricultural yield prediction. While not a replacement for agronomic models or remote sensing approaches, the Framework 2 Model 1 results provide evidence that hydrological and nutrient-driven environmental variability plays a measurable role in shaping interannual agricultural performance.

5.3.2 Pine Barrens

Incorporating groundwater and contaminant data from the Long Island Pine Barrens into the yield model provides a distinct environmental perspective. The Pine Barrens represent a major recharge zone with detailed monitoring of nutrients, inorganics, pharmaceuticals, pesticides, and groundwater levels. Although these signals do not map directly to the agricultural counties included in yield prediction, they serve as indicators of broader hydrological and subsurface conditions that co-vary with regional climate and environmental stress.

The Random Forest classifier trained on the Pine-derived feature set achieved a macro-averaged F1 score of 0.395 (± 0.109) and a macro ROC AUC of 0.669 (± 0.059). This performance is modestly lower than the surface-water-based model, which is expected given that Pine Barrens data capture only three years of measurements and reflect localized aquifer conditions rather than statewide agricultural environments. Nevertheless, the model significantly outperforms the majority-class baseline, indicating the presence of useful environmental signal even under these constraints.

SHAP-based interpretation provides insight into how the high-dimensional Pine dataset contributes to yield prediction. Several groundwater-level indicators, specific dissolved constituents, and select pharmaceutical or pesticide detections emerge as influential for distinguishing high-yield years. These variables likely function as proxies for climatic wetness, hydrological recharge, or anthropogenic pressure, allowing the model to indirectly infer environmental conditions relevant to crop growth. The SHAP global importance profile shows a diverse set of drivers, reflecting the high dimensionality and multicollinearity characteristic of groundwater chemical datasets.



Metric	Random Forest (Yield + Pine)	Dummy baseline
Macro F1 (mean ± std)	0.395 ± 0.109	0.194 ± 0.006
Macro AUC (mean ± std)	0.669 ± 0.059	0.500 ± 0.000

Figure 9 – Training Results for Crop Yield Prediction Framework: Groundwater Chemistry

Despite the limited temporal overlap between Pine Barrens monitoring and the yield dataset, the model demonstrates that subsurface environmental conditions, even when geographically distant, capture components of broader environmental variability. This finding highlights the potential of integrating nontraditional hydrological datasets into agricultural prediction frameworks, especially when paired with interpretable machine-learning tools that can disentangle complex signal sources.

6. Discussion

6.1 Stronger Predictability in Surface Waters

A central finding of this study is the markedly stronger predictability of microbial events in surface waters compared to the distribution system. While the NYC distribution network model performed substantially better than the majority-class baseline, the predictive signal was inherently limited by the rarity of events and the constrained variability of a highly regulated system. Surface waters, by contrast, exhibited rich environmental gradients – particularly in nutrient concentrations, turbidity, conductivity, and temperature – that translated into clearer and more informative patterns of microbial exceedance. The surface-water model's ROC AUC near 0.90 and PR AUC above 0.67 demonstrate that routine physicochemical measurements capture robust signatures of microbial risk at the monthly scale.

This difference reflects underlying system characteristics. Drinking-water distribution networks are designed for stability; operational controls maintain narrow ranges of disinfectant and physical parameters. Consequently, microbial detections are episodic and often linked to localized failures or rare hydraulic anomalies, reducing the extent to which monthly averages can explain event occurrence. In contrast, surface waters are inherently dynamic systems influenced by watershed processes, weather patterns, land-use activities, and point and nonpoint pollution sources. These variables collectively shape nutrient loading, particle concentration, and microbial transport, producing strong, persistent associations that are detectable through aggregated monthly monitoring.

The dominance of nitrate and phosphorus as predictors emphasizes the role of eutrophication pathways in governing microbial conditions across the state. These nutrients stimulate primary production and microbial metabolism directly while also signaling the presence of agricultural and wastewater runoff. Turbidity further enhances microbial survival and transport, acting as both a physical shield and a proxy for erosion or storm-driven pulses. The consistent prominence of these variables across all cross-validation folds reinforces the structural strength of the predictive relationships.

Taken together, the results demonstrate that surface-water microbial exceedances are not random occurrences but are fundamentally shaped by chemical and hydrological regimes that can be captured with surprisingly high fidelity using standard monitoring programs. This has important implications for statewide water-quality forecasting, early warning systems, and watershed management strategies that emphasize nutrient reduction and stormwater control.

6.2 Insights for Agricultural Productivity

The agricultural yield models reveal that environmental indicators extracted from water-quality datasets carry non-trivial predictive signal for county-level crop performance. Although the models are not agricultural forecasting tools per se, they illuminate meaningful connections between hydrological variability and agricultural outcomes. The surface-water model (Framework 2 Model 1), in particular, shows that yield classes correlate with patterns of turbidity, nitrate, conductivity, and microbial stress at the statewide scale. These associations suggest that hydrologic conditions, such as rainfall intensity, runoff, nutrient mobilization, and watershed wetness, serve as broad indicators of growing-season constraints experienced across agricultural regions.

Turbidity and microbial event rates emerge as especially informative predictors. Elevated turbidity in surface waters often corresponds to years with heavier rainfall or soil disturbance, conditions that can impede field operations, reduce soil aeration, or promote plant disease. Similarly, high microbial event rates indicate periods of hydrological disturbance or nutrient flux, which frequently coincide with stresses such as waterlogging or nutrient imbalance in agricultural soils. Thus, even though these water-quality indicators do not directly affect crops, they reflect systemic environmental pressures that shape crop development at scale.

The Pine Barrens–based model extends this perspective by demonstrating that groundwater and contaminant signals, even from a localized aquifer system, exhibit detectable associations with yield patterns. These relationships likely arise from shared climatic drivers. For example, groundwater levels respond to precipitation anomalies, which in turn affect soil moisture availability, planting windows, and drought risk. The presence of pharmaceuticals or pesticides in groundwater may reflect regional patterns of wastewater infiltration or land-use intensity, indirectly signaling broader anthropogenic pressures that co-vary with agricultural stress.

These findings underscore an emerging insight: water-quality datasets, when aggregated appropriately, encode information about the environmental regime that crops experience, even when derived from monitoring networks not originally designed for agricultural applications. This opens the door to the development of multi-sector environmental intelligence tools capable of integrating hydrological, chemical, and ecological signals to contextualize agricultural trends.

6.3 Limitations

Several limitations temper the interpretation of these findings. First, class imbalance significantly influenced the drinking-water model. Microbial detections in distribution systems are extremely rare, limiting the sensitivity of monthly-scale predictive models and constraining the PR AUC achievable under such conditions. Second, surface-water datasets, while extensive, are unevenly distributed across geography, monitoring frequency, and analyte completeness. Not all stations sample all variables at consistent intervals, and nutrient or microbial records may be missing in specific years or regions. Although aggregation reduces noise, it cannot overcome fundamental gaps in temporal coverage.

Third, the agricultural yield models rely on statewide or regionally averaged water-quality indicators, which inevitably obscure finer-scale environmental heterogeneity. County-level yields integrate the effects of management practices, soil quality, crop genetics, pest pressure, and climate variability, factors not captured by water datasets alone. Thus, the predictive performance should be interpreted as capturing broad environmental alignment rather than causal mechanisms.

Finally, the Pine Barrens dataset spans only three years, limiting the temporal overlap with yield records and constraining the stability of the high-dimensional model. The strong SHAP signals must therefore be viewed as exploratory rather than definitive. Groundwater signals from one aquifer region cannot be assumed to generalize statewide, and more localized agricultural–hydrological data integration would be needed to validate these relationships.

6.4 Future Work

Further research should explore several key directions. One priority is the incorporation of higher-frequency water-quality and hydrological data, such as real-time turbidity, discharge, and rainfall measurements, which may substantially improve predictive performance for microbial events and agricultural yields. Integrating these datasets into a unified spatiotemporal graph or sequence model, such as an LSTM or transformer architecture, may capture temporal dependencies that static monthly features cannot.

A second avenue involves expanding the agricultural modeling component by integrating remote sensing indicators, gridded climate datasets, and soil moisture products. These datasets would help disentangle the roles of hydrology, climate, and management in determining yields and would allow water-quality variables to function as complementary environmental indicators rather than primary predictors.

Third, the groundwater analysis should be expanded beyond the Pine Barrens region. Linking agricultural counties to their own aquifer systems, using USGS groundwater networks, local monitoring wells, or state environmental data, would enable far more geographically relevant inference. A statewide groundwater–surface-water–yield fusion model may reveal new interactions between subsurface storage dynamics, watershed runoff, and agricultural performance.

Finally, the framework could be extended into a statewide water–agriculture early warning system, using predictive modeling to identify years at heightened risk for microbial exceedances, nutrient stress, or yield reductions. Such a system would support water managers, agricultural agencies, and policymakers seeking integrated environmental intelligence for climate resilience, watershed protection, and food-system planning.

7. Conclusion

This study demonstrates that integrating drinking-water, surface-water, groundwater, and agricultural datasets into a unified spatiotemporal machine-learning framework offers meaningful insights into the relationships between water quality and agricultural productivity in New York State. Despite their differing purposes and sampling regimes, these datasets contain overlapping environmental signals that, when harmonized and aggregated appropriately, reveal strong predictive structure. Surface waters exhibit clear and robust relationships between nutrient enrichment, turbidity, and microbial exceedances, and these relationships manifest in high model performance. Distribution systems, though more constrained and operationally stable, still show detectable patterns linking disinfectant and particle metrics to microbial anomalies.

Agricultural yields, meanwhile, exhibit measurable associations with annual hydrological and water-quality conditions. While not deterministic, these relationships indicate that water-quality datasets can serve as valuable environmental proxies that reflect broad-scale climatic and watershed processes influencing crop outcomes. The extension into groundwater chemistry further illustrates the potential of incorporating nontraditional environmental indicators into agricultural analytics, even when temporal overlap is limited.

Collectively, the results present a promising direction for multi-dataset environmental intelligence, in which water systems and agricultural systems are analyzed not as isolated domains but as interconnected components of a shared hydrological and ecological landscape. The modeling framework offers a scalable foundation for linking monitoring networks across sectors and for developing predictive tools that can inform water management, agricultural planning, and climate adaptation strategies statewide.

References

ScienceBase Instructions and Documentation. "API and Web Services." n.d. *USGS* . URL: https://www.usgs.gov/sciencebase-instructions-and-documentation/api-and-web-services. 2025.

Department of Environmental Conservation. *Groundwater Study at Long Island Mines Work Plan*. Kathy Hochul, 2021.

Karpatne, Anuj, et al. "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data." 27 Dec 2016. DOI: https://doi.org/10.1109/TKDE.2017.2720168.

Kulyal, Malika and Parul Saxena. "Machine Learning approaches for Crop Yield Prediction: A Review." *7th International Conference on Computing, Communication and Security*. Seoul: IEEE Xplore, 2022. DOI: 10.1109/ICCCS55188.2022.

Marrone, Babetta L., et al. "Toward a Predictive Understanding of Cyanobacterial Harmful Algal Blooms through AI Integration of Physical, Chemical, and Biological Data." *ACS EST Water* (2023): 844-858. DOI: https://doi.org/10.1021/acsestwater.3c00369.

Mukti, Ishrat Zahan and Dipayan Biswas. "Transfer Learning Based Plant Diseases Detection Using ResNet50." *4th International Conference on Electrical Information and Communication Technology (EICT)*. Khulna, Bangladesh, 2019.

Rule, Adam, et al. "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks." *PLOS Computational Biology* (July 25, 2019). https://doi.org/10.1371/journal.pcbi.1007007 .

Salami, Zayd Ajzan, et al. "Modelling Crop Yield Prediction with Random Forest and Remote Sensing Data." *Natural and Engineering Sciences* (2025): 67-78. DOI: 10.28978/nesciences.1763843 .

Vermont Department of Health. "Tracking Cyanobacteria in Vermont." n.d. *Vermont Department of Health.* URL: https://www.healthvermont.gov/environment/tracking/cyanobacteria-blue-green-algae-tracker.

Wickham, Hadley. "Tidy Data." *Jouernal of Statistical Software* (Sep 12, 2014). DOI: 10.18637/jss.v059.i10.

"*2022–2023 Harbor Survey Report*." *New York City Department of Environmental Protection*, 2024. https://storymaps.arcgis.com/stories/12b3077dd5c84f2b81f81ea4bab07b18

"*5 Reasons Why Water Quality Analysis Is Important for the Chemical Manufacturing Industry.*" Meena Sankaran https://ketos.co/5-reasons-why-water-quality-analysis-is-important-for-the-chemical-manufacturing-industry

"*Characteristics of Water: Physical, Chemical, and Biological Aspects*." *Dairy Equipment & Utilities*, 5 Feb. 2024, https://agriculture.institute/diary-equipment-utilities/characteristics-of-water-physical-chemical-biological/.

"*The Essential Role of Chemicals in Water Treatment.*" Reachem, August 20, 2024 https://www.reachemchemicals.com/blog/the-essential-role-of-chemicals-in-water-treatment/

"*Three Main Types of Water Quality Parameters Explained*." *Sensorex*, 20 Sept. 2021, https://sensorex.com/three-main-types-of-water-quality-parameters-explained/