Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning and Citizen Science for Catchment-Scale Water-Quality Assessment

Xinyu Liu
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Heri Chisute
Ministry of Water, Dodoma, Tanzania

Fred Nyongesa
Water Resources Authority (WRA), Nairobi Kenya

Rochi Mkole
Ministry of Water, Musoma, Tanzania

Stuart Warner
Global Environment Monitoring Unit, United Nations Environment Programme (UNEP), Nairobi, Kenya

Georgio Emmanuel Nader
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Amedeo Boldrini
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Alessio Polvani
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Riccardo Cirrone
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Luisa Galgani
University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Steven Loiselle
steven.loiselle@unisi.it

University of Siena Department of Biotechnology Chemistry and Pharmacy: Universita degli Studi di Siena Dipartimento di Biotecnologie Chimica e Farmacia

Research Article

Keywords: Environmental pollution monitoring, Water quality assessment, Citizen science, Interpretable machine learning, Transboundary river basin monitoring

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10575670/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Machine Learning and Citizen Science for Catchment-Scale Water-Quality Assessment

Xinyu Liu [a], Heri Chisute [d], Fred Nyongesa [h], Rochi Mkole [g], Stuart Warner[f], Georgio Emmanuel Nader [a], Amedeo Boldrini [a], Alessio Polvani [a], Riccardo Cirrone [a, c], Luisa Galgani [a, e], Steven Loiselle [a, b, e *]

[a] *Department of Biotechnology Chemistry and Pharmacy, The University of Siena, Siena, 53100, Italy*
[b] *Earthwatch Europe, 102-104 St Aldate's, Oxford, OX1 1BT, United Kingdom*
[c] *Department of Earth and Marine Sciences, University of Palermo, Via Archirafi, 22 90123 Palermo -Italy*
[d] *Ministry of Water, Dodoma, Tanzania*
[e] *Center for Colloids and Surface Science (CSGI)-Siena Research Group, University of Florence, Via della Lastruccia 3, 50019 Firenze, Italy*
[f] *Global Environment Monitoring Unit, United Nations Environment Programme (UNEP), Nairobi, Kenya*
[g] *Ministry of Water, Musoma, Tanzania*
[h] *Water Resources Authority (WRA), Nairobi Kenya*

*Corresponding authors:
Addresses: *Department of Biotechnology Chemistry and Pharmacy, University of Siena, via Aldo Moro, 2, 53100, Siena, Italy*; E-mail addresses: steven.loiselle@unisi.it (S.Loiselle*).

E-mail addresses: x.liu@student.unisi.it (X. Liu), heri.chisute@maji.go.tz (H. Chisute), fred.nyongesa@wra.go.ke (F. Nyongesa), rochimkole56@gmail.com (R. Mkole), stuart.warner@un.org (S. Warner), g.nader@student.unisi.it (G.Nader), amedeo.boldrini@student.unisi.it (A.Boldrini), alessio.polvani@student.unisi.it (A.Polvani), riccardogaetano.cirrone@unipa.it (R.Cirrone), luisa.galgani@unisi.it (L.Galgani), steven.loiselle@unisi.it (S.Loiselle).

Abstract
Assessing and managing water quality in data-scarce tropical basins is increasingly challenging under accelerating land-use change, intensifying human pressures, changing livelihood practices, and climate driven shifts in precipitation regimes. In the transboundary Mara River Basin (MRB), these pressures strongly influence sediment and nutrient dynamics, yet conventional monitoring networks operated by national authorities lack the temporal and spatial resolution required to characterize pollution sources and event-driven responses. To address these gaps, this study integrates citizen-science observations, with known and quantified uncertainty characteristics, with satellite derived hydro-climatic and land-use variables within an interpretable machine learning framework to investigate the drivers of turbidity (NTU), nitrate (NO_3), and phosphate (PO_4) dynamics. Citizen scientists were recruited, trained, and equipped by local ministry staff following standard methods and measurements were quality controlled before analysis. Random Forests (RF) and Artificial Neural Networks (ANN) were developed to predict water-quality variability and quantify the relative influence of environmental and anthropogenic drivers. RF provided consistently better predictive performance than ANN, indicating that tree-based ensemble methods are well suited to heterogeneous citizen-science datasets while providing interpretable estimates of predictor importance. Model interpretation using SHAP values revealed that precipitation and river flow velocity dominate short-term, event-based fluctuations of turbidity, while population density was an important predictor of NO_3, consistent with diffuse sources. Site-level analyses further demonstrated pronounced spatial heterogeneity in precipitation sensitivity, identifying mid- to lower-basin reaches as particularly responsive to rainfall-driven sediment mobilization. These findings demonstrate that integrating citizen science with interpretable machine learning can strengthen pollution assessment, identify catchment-scale drivers and hotspots, and support adaptive water-quality management in data-scarce river basins undergoing rapid land-use and climate change.

Keywords: Environmental pollution monitoring; Water quality assessment; Citizen science; Interpretable machine learning; Transboundary river basin monitoring.

1.Introduction

Freshwater ecosystems worldwide are experiencing increasing pressure from rapid population growth, land‐use change, agricultural practices, urban expansion, and climate variability. These drivers collectively contribute to the deterioration of riverine water quality, manifested through elevated turbidity, nutrient enrichment, organic pollution, and altered runoff dynamics, threatening the stability and resilience of freshwater systems (Vörösmarty et al., 2010; Meybeck, 2003; Odero, 2024; Parmar & Bhardwaj, 2015). In many developing regions, water‐quality monitoring networks remain incomplete or spatially sparse due to financial and logistical constraints, making it difficult to accurately assess temporal and longitudinal trends in water quality (UNEP, 2016; Palmer et al., 2008; UNEP, 2024). These challenges are further amplified in transboundary rivers, where differences in monitoring systems, national priorities, and resource allocation introduce additional uncertainties to collaborative water governance (McClain et al., 2014).

Transboundary river systems in East Africa represent some of the most complex and challenging freshwater environments on the continent. Among them, the Mara River Basin (MRB) serves as a critical ecological and socio-economic lifeline, supplying domestic, agricultural, and pastoral water to more than 1.1 million people, while traversing globally significant ecological landscapes such as the Maasai Mara National Reserve and Serengeti National Park (Kihwele, 2021; Komba, 2022). The river plays a fundamental role in sustaining the large migratory wildlife populations that depend on these ecosystems (John et al., 2018). However, rapid indigenous vegetation clearing, agricultural intensification and changing land use practices, population pressure, expansion of livestock density, and the proliferation of tourism infrastructure have imposed increasing pressure and conflicts on water resources across the basin (Kihwele, 2021; Mango et al., 2011a; Mango et al., 2011b). These pressures are further intensified by climate variability, which amplify hydrological extremes and alter water-quality responses across the basin.

In addition to these anthropogenic pressures, climate variability and ongoing climate change exert strong controls on water-quality dynamics in the Mara River Basin, making them central to this study. Increasingly erratic rainfall patterns, shifts in seasonal timing, and the intensification of extreme storm events across East Africa (Niang et al., 2014; Rowell et al., 2015) compounded by land-use practices directly influence sediment mobilisation, nutrient flushing, and pollutant dilution. Heavy rainfall episodes rapidly accelerate soil erosion and runoff, generating sharp increases in turbidity and nutrient loads, while prolonged dry periods tend to concentrate pollutants and heighten the impacts of point-source discharges. These hydrological fluctuations complicate efforts to identify pollution drivers and to interpret short-term water-quality responses, especially in data-scarce regions reliant on citizen-science monitoring. As climate change continues to amplify rainfall extremes, understanding how hydroclimatic variability interacts with land-use pressures becomes essential for improving predictive modelling and prioritising management interventions in the MRB.

These pressures are common to tropical rivers, where the intrinsic complexity of water-quality dynamics is compounded by variable seasonal rainfall patterns leading to hydrological variability (Bewket & Sterk, 2005). Processes such as soil erosion, sediment transport, and nutrient loading are often event-driven and characterized by short-lived but intense pulses (Bewket & Sterk, 2005; Dunne, 1979). Coupled with limited monitoring infrastructure and sparse long-term datasets, management authorities struggle to manage such dynamic and heterogeneous changes. Consequently, alternative and complementary monitoring approaches are urgently needed to support effective water-resource management and ecological conservation (Kihwele, 2021; Komba, 2022). This is particularly critical in areas where declining water quality for domestic use necessitates capital-intensive conventional water treatment (Soetan et al., 2024).

Against this backdrop, citizen science has emerged as a valuable supplementary data source for environmental monitoring for agencies and national ministries in Africa (Moshi et al., 2023; Taylor et al., 2022). Through the establishment of extended monitoring network based on a consistent and quality-controlled monitoring approach, new insights on pollution dynamics and hotspots can be identified (Buytaert et al., 2014; Loiselle et al., 2024). These efforts substantially expand the spatial coverage of monitoring networks and provide systematic observations in areas that were previously difficult to access. Owing to its relatively high sampling frequency and diverse data types, citizen-science monitoring offers unique advantages for interpreting short-term water-quality fluctuations and identifying potential pollution events. Interestingly, despite the constraints, local communities have been shown to embrace citizen science through information sharing and the establishment of water quality monitoring station (Moshi et al., 2023).

However, citizen-science water-quality data often display pronounced skewness and heterogeneity due to comparatively high limits of quantification, coarser resolution (e.g., ordinal classes), and operator subjectivity (visual interpretation, lighting). These characteristics often lead to non-normally distributed, zero-inflated, and noisy data, limiting the use of analytical approaches that assume linearity, homoscedasticity, and normal error structures (Nokelainen et al., 2024). Consequently, methods capable of handling nonlinear responses and robust to noisy, heterogeneous, and partially uncertain data inputs, are particularly suitable for citizen-science datasets. In recent years, machine-learning methods have gained increasing attention in water-quality assessment and prediction due to their capacity to model nonlinear relationships, integrate multidimensional predictors, and remain robust in the presence of noisy, skewed, and censored data (Zounemat-Kermani et al., 2021).

Random Forests (RF) and Artificial Neural Networks (ANN) represent two widely used and conceptually distinct model families, each offering unique structural advantages and application potentials. RF, an ensemble method based on multiple decision trees, reduces model variance through bootstrap sampling and aggregated voting, enabling robust performance even under sparse, irregular, or noise-dominated conditions. In addition, RF provides feature-importance measures that facilitate the interpretation of how environmental variables contribute to water-quality variations, making it a popular choice in hydrological and environmental research (Bargam et al., 2024; Nishat et al., 2025). In contrast, ANN exhibits stronger nonlinear fitting capabilities and can capture complex eco-hydrological interactions through multi-layer network structures. Its sensitivity to seasonal dynamics and ability to capture latent relationships among variables makes ANN particularly suitable for water-quality indicators governed by intricate or non-linear patterns (Pan et al., 2025; Lokman et al., 2025). Despite the potential of both models, there remains limited work explicitly integrating citizen-science water-quality data with interpretable machine-learning approaches to (i) identify sub-basin-scale pollution dynamics, (ii) quantify spatial heterogeneity in hydroclimatic sensitivity, and (iii) translate model outputs into decision-relevant insights for basin-scale management in data-scarce regions.

The main objectives of this study are (1) integrate citizen-science and Earth observation datasets within an interpretable machine-learning framework to characterize spatial and temporal water-quality patterns; (2) identify and compare dominant environmental and anthropogenic drivers across sub-basins; (3) assess the spatial variability in precipitation sensitivity with respect to changes in different water-quality variables; and (4) evaluate the consistency and explanatory capacity of different machine-learning approaches for decision-relevant insights in data-scarce river basins.

2. Methods

The Mara River is one of the most ecological and socioeconomically significant transboundary river systems in East Africa. Originating in the Mau Forest Complex in the highlands of southwestern Kenya, with main tributaries as Nyangores, Amala, Talek and Engare, the river flows for approximately 395 km within the savannah grasslands of Masai Mara National Reserve before entering northern Tanzania, where it traverses the Serengeti National Park and ultimately drains into Lake Victoria (Figure 1). The basin covers an area of approximately 13,500 km², with roughly two-thirds located in Kenya and one-third in Tanzania. In recent decades, increasing pressures from small- and large-scale agriculture (vegetation clearing), earth-road networks, livestock production, and tourism, combined with prolonged droughts linked to climate variability, have contributed to reduced downstream flows and increased sediment and nutrient delivery to Lake Victoria (Nyairo et al., 2015; McClain et al., 2014). Current estimates indicate that 35–47 million people reside within the Lake Victoria basin, with an average population density of approximately 250 people per km², making it one of the most densely populated inland regions in Africa (Odenyo, 2018; Wanjiku & Kennedy-Walker, 2024). Livelihoods in many areas remain heavily dependent on fisheries, smallholder agriculture, and informal economic activities, while continued population growth and climate change have increased vulnerability to water scarcity, flood hazards, pollution and ecosystem degradation (Nyamweya et al., 2020).

Across the Mara River Basin, monitoring sites (40) were established to capture major tributaries and key locations along the main river stem (Figure 1). Site selection aimed to characterize baseline conditions, complement ongoing national monitoring programs, and target areas with potential pollution pressures. In both Kenya and Tanzania, water resources are managed locally by self-driven entities formed and registered in accordance with national regulations for conservation, protection and advocacy for sustainable water use; Water Resource Users Associations (WRUAs) in Kenya and Water Resource Associations (WRAs) in Tanzania. Within this framework, five monitoring sites in Kenya are coordinated by the Water Resources Authority and seven sites in Tanzania are monitored by the Ministry

of Water. Citizen scientists were recruited from WRUA/WRA membership and trained to conduct regular monthly monitoring of chemical, optical, ecological, and hydrological variables (Bishop et al., 2025). Monitoring began in February 2023 and is ongoing, with data used in the present study extending through November 2025. The monitoring network expanded in June 2024. Only sites with a minimum of eight observations were included in this analysis, resulting in a median of 22.5 measurements per site and a total of 945 individual datasets.

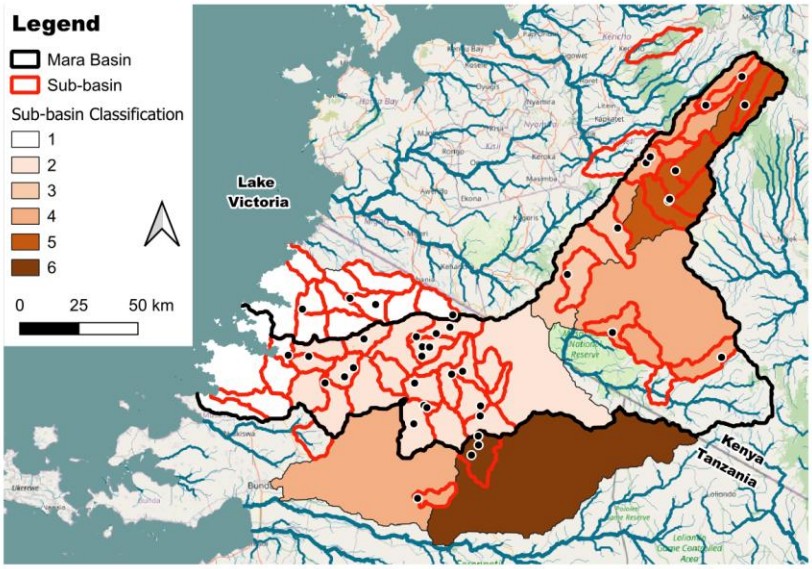


Figure 1: Monitoring network and sub-basin delineation in the Mara River basin. Black dots represent monitoring stations, red lines indicate sub-basin boundaries, and the black outline is the basin's extent. Brown shades indicate the upper basin–lower basin sequence (dark (eg. 6)= upstream; light = downstream (eg. 1)).

Trained citizen scientists measured turbidity (NTU), nitrate (NO_3), and phosphate (PO_4), as well as local ecological conditions at each site. Measurements were uploaded directly to a project platform, where Ministry and Agency scientists, together with Earthwatch scientists, conducted regular validation procedures (Bishop et al., 2025). All three water-quality variables exhibited positively skewed distributions. However, these datasets also contain inherent uncertainties associated with ordinal classification, detection limits, and observer interpretation, which may introduce measurement noise and reduce precision at fine scales. Citizen scientists also recorded dominant land use within 100 m of each site using categorical classifications. Streamflow velocity was recorded in four relative velocity classes selected from a dropdown menu, enabling comparative analysis across monitoring events. The reliability of citizen-science measurements used in this study has been previously evaluated under comparable field conditions (Moshi et al., 2023). In that study, nitrate measurements showed an overall agreement of 91% with professional measurements (Cohen's $\kappa = 0.74$, $p < 0.001$). Phosphate measurements showed an agreement of 74% (Cohen's $\kappa = 0.61$, $p < 0.001$), with 96% of discrepancies falling within one concentration class. Turbidity measurements derived from Secchi tube observations demonstrated strong agreement with laboratory-derived total suspended solids (TSS), with correlation coefficients of 0.79 ($p < 0.001$) and classification accuracy exceeding 85% across sites. These results indicate that, while subject to observer variability and classification uncertainty, citizen-science measurements provide consistent and environmentally meaningful signals suitable for basin-scale analysis.

To explore environmental and anthropogenic characteristics of the monitoring network, global geospatial datasets partitioned into discrete hydrological sub-basins were used, primarily derived from HydroSHEDS BasinATLAS (Linke et al., 2019). Predictor variables included both dynamic factors (e.g., monthly precipitation and temperature) and static or slowly varying variables (e.g., land use, forest cover, soil erosion, subbasin order and human footprint indices). While dynamic variables capture short-term temporal variability, static variables represent the underlying spatial characteristics of each sub-basin and provide contextual controls on environmental processes. From these datasets, a suite of indicators was selected to characterize different dimensions of watershed conditions. These included measures of population pressure (population density and total population), land-use intensity (cropland extent,

pasture extent, and urban area), land-cover conditions (forest cover, soil erosion levels, and dominant land-cover types), and aggregated anthropogenic activity (night-time light intensity, road density, and human footprint indices). Soil erosion values from data HydroSHEDS BasinATLAS, were previously estimated based on an improved large-scale version of the Revised Universal Soil Loss Equation (RUSLE) model. Together, these variables capture both natural and human-driven influences on water quality across the basin. Data were available at both the sub-basin level and as upstream-aggregated values. In this study, sub-basin-level data were extracted for each monitoring location to represent local environmental conditions. Monthly averaged climate data, including total precipitation and mean air temperature at 2 m, were obtained from the Copernicus Climate Change Service ERA5 dataset (Hersbach et al., 2023) via the Climate Data Store (CDS), ensuring temporal consistency with the monthly water-quality observations. Environmental predictor variables were therefore derived at different spatial and temporal scales depending on their characteristics. Land-use and anthropogenic variables (e.g., population density, cropland extent, pasture extent, forest cover, and human footprint) were calculated at the sub-basin level and treated as static or slowly varying over the study period, representing local baseline conditions associated with each monitoring site. In contrast, climate variables (e.g., precipitation and temperature) were included as time-varying predictors to capture short-term environmental dynamics.

Several ANN and RF based models were selected to capture nonlinear interactions between environmental and anthropogenic variables after exploratory attempts using linear and probabilistic approaches, including linear regression, discriminant analysis, and logistic regression, which did not provide statistically significant results. To facilitate comparison of modelling approaches, SHapley Additive exPlanations (SHAP) were used to quantify the contribution of the 13 environmental and anthropogenic variables to the model for each water quality variable (Lundberg & Lee, 2017; Lipton, 2018). SHAP enables interpretation of machine-learning predictions by quantifying the relative contribution of input features to model outputs, thereby identifying statistical associations between predictors and predicted water-quality responses (Yuan et al., 2025; Gao et al., 2025). It is important to note that SHAP values do not imply causal relationships but rather reflect how the model attributes variation in predictions to input features. The combined use of ANN and RF approaches enabled methodological triangulation and strengthened confidence in inferred driver–response relationships (Li et al., 2019). Furthermore, Spearman correlation coefficients were calculated between each predictor variable and the observed water quality indicators (NTU, NO_3, and PO_4) across all sites–time observations. These were used to indicate the monotonic direction (positive or negative) of association between predictors and response variables, independent of the machine learning model.

A range of ANN architectures was evaluated to assess suitability given the size and complexity of the dataset (Supplementary Table 1). A sequential feedforward ANN was selected due to its stability, reduced risk of overfitting, compatibility with moderate-sized datasets, and ability to model nonlinear environmental processes in environmental data (Goodfellow et al., 2016; Haykin, 2009). Environmental predictors (e.g., precipitation, land-cover composition, soil erosion, anthropogenic variables, and derived sub-basin order), together with temporally derived features (e.g., lagged values and rolling averages of the target variable and precipitation), were supplied to the input layer. To explicitly account for spatial heterogeneity, monitoring sites were encoded using one-hot representations, allowing the model to capture site-specific characteristics within a unified spatiotemporal framework. The ANN architecture consisted of two fully connected hidden layers with rectified linear unit (ReLU) activation functions, followed by a linear output layer producing continuous predictions for each water-quality indicator (NTU, NO_3, or PO_4). Dropout regularization was optionally applied between layers to mitigate overfitting. Model training minimized mean squared error (MSE) using the Adam optimizer with adaptive learning rates and backpropagation. To preserve temporal integrity and avoid information leakage, the dataset was split chronologically, with the earliest 75% of observations used for training and the most recent 25% reserved for independent testing. All input features were standardized using Z-score normalization based solely on the training subset, and the same transformation was applied to the test data. Model hyperparameters, including hidden layer sizes, dropout rates, learning rates, and batch sizes—were systematically optimized through a grid search procedure combined with early stopping based on validation loss. The optimal model configuration was selected according to performance on the independent test dataset. To ensure comparability with the RF model and to facilitate interpretation in real-world units, model performance was evaluated using R^2, MAE, and RMSE calculated on back-transformed predictions (i.e., after reversing the log transformation applied to the target variable). Following model training, predictions were generated across the full spatiotemporal domain for subsequent analysis.

Several Random Forest (RF) regression models were developed using ensembles of decision trees trained on bootstrap samples, with predictions obtained as the ensemble mean (Breiman, 2001; Zounemat-Kermani et al., 2021). The modelling framework adopted an explicit spatiotemporal structure in which each observation corresponds to a specific monitoring site at a given time step (site × time). This design enables the model to simultaneously capture spatial heterogeneity across monitoring locations and temporal dynamics within individual sites. Models were constructed to predict monthly NTU, NO_3, and PO_4 concentrations using pooled observations across all sites while strictly preserving temporal ordering. This time-aware structure was implemented to prevent information leakage and reduce overfitting associated with temporal autocorrelation in environmental datasets. Specifically, a forward-chaining validation strategy was applied, in which the earliest 75% of observations were used for model training and the most recent 25% were reserved as an independent test set. To explicitly account for temporal dependencies, lagged predictors (e.g., precipitation at multiple time steps) and rolling-window statistics (e.g., 3-month and 6-month moving averages) were incorporated. These features allow the model to capture both short-term memory effects and smoothed temporal trends in water-quality dynamics. Spatial structure was represented using both categorical site-level encoding and a derived continuous variable, which reflects the relative upstream–downstream position of each monitoring site within the river network. Together, these variables provide a simplified but effective representation of spatial organization and hydrological connectivity.

Hyperparameter optimization was performed using grid search in combination with time-series cross-validation (TimeSeriesSplit, k = 5) (75% training part), ensuring that validation folds respected temporal ordering. The parameter space included the number of trees (n_estimators), maximum tree depth (max_depth), minimum samples per leaf (min_samples_leaf), minimum samples required for splitting (min_samples_split), and the number of features considered at each split (max_features). The optimal parameter combination was selected based on cross-validated R^2 scores. A fixed random state was used to ensure reproducibility. Unlike artificial neural networks, RF models do not require feature scaling and are robust to non-linear relationships and interactions among predictors. Model performance was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE), calculated on the independent test dataset in the original (back-transformed) units of the response variables. Following model evaluation, predictions were generated across the full spatiotemporal dataset. Basin-averaged time series were further derived by aggregating predictions across all monitoring sites for each time step, providing a basin-scale representation of temporal dynamics while acknowledging that such aggregation may smooth local variability and extreme events.

It should be noted that the model includes lagged variables when making predictions, with the aim of maximizing predictive performance; however, these lagged variables were excluded for the SHAP analysis to isolate the effects of environmental drivers.

All analyses were conducted using Python, with TensorFlow, Scikit-learn, SHAP, and Pandas libraries.

3. Results

3.1 Spatial heterogeneity of water quality and land use features across the Mara Basin

NTU, NO_3, and PO_4 exhibit pronounced spatial heterogeneity across the study watershed (Supplementary Table 2). Turbidity (NTU) medians across the 40 monitoring sites ranged from < 14 NTU at Bwawa La Manchira intake to 240 NTU at numerous sites (e.g., Mto Mara Kivuko), reflecting local and cumulative land-use pressures. Moderate levels dominated mid-basin sites like Nyangores River at Bomet Town (40 NTU, n=33), while upper tributaries showed cleaner waters (e.g., Upper Nyangores at TMC bridge, 23 NTU, n=10). Sites in the lower basin had typically elevated turbidity (e.g. Mto Tigite Juu, 75 NTU, n=37) underscoring the value of longitudinal CS monitoring for adaptive management.

NO_3 medians varied widely from low concentrations (0.1 mg/L) in both upper and lower catchment areas (Bwawa La Manchira and upper Nyangores) and high concentrations (>1 mg/L) in middle and lower catchment sites (Mto Mori - keisangora, Isei river bridge)- Most sites (over 70%) clustered around 0.35-0.75 mg/L.

PO_4 medians were generally low (<0.1 mg/L) across sites, with the majority at 0.035 mg/L and minima of 0.01 mg/L in the upper catchment (Bwawa La Manchira and upper Nyangores). Elevated medians occurred in several sites in the middle catchment: Marasienta - Mara River (0.15 mg/L), Mto Mara Nyamerambaro Kiriba (0.2125 mg/L), and Mto Mori -keisangora (0.15 mg/L), suggesting localized hotspots of phosphorus loading within the watershed.

The 32 sub-basins containing the 40-monitoring location had highly different features (Table 1). Sub-basin areas were smallest in upper headwater catchments of the Nyangores and Amala Rivers (Kenya). Forest and pasture extent were similar between Kenyan and Tanzania sub-basins, while cropland extent was highest in the lower basins (Tanzania). These data indicate that erosion rates were highest in the Tanzanian sub-basins. Population features were similar over the macro-areas, but very different between sub-basins, as were related features such urban extent, road density and night-time lights. Median Human footprint values for Tanzania and Kenya showed different results for many sub-basins between 1993 and 2009. In 1993, Kenyan sub-basins had statistically high footprint values, which was not the case for 2009, indicating an increasing pressure in Tanzanian sub-basins. To capture this, the 1993-2009 change in Human footprint was used as a proxy for land use change, and the 1993 footprint layer was dropped.

Table 1. Sub-basin features, minimum, median and maximum values

Sub-basin feature	Units	min	median	max
precipitation	millimeter/month	21.7	155	244
forest percent of sub-basin	percent cover	0.0	36	100
cropland extent of sub-basin	percent cover	3.0	10	28
Pasture extent of sub-basin	percent cover	1	41	93
Soil Erosion of sub-basin	kg/hectare per year	2600	11301	29390
Population count of sub-basin	people count	1579	21376	64328
Population density of sub-basin	people per km2	8912.0	106455	449003
Urban extent of sub-basin	percent cover	0.0	2	78
Night-time lights of sub-basin	index value x100	0.0	23	988
Road density of sub-basin	meters per km2	4	103	580
Human footprint of sub-basin 2009	index value x10	45	108	200
Change in Human footprint of sub-basin 1993 to 2009	index value x10	-42	11	30
River surface velocity (CS observed)	Ordinal	≈ 0 m/s 1		>1.0 m/s

3.2 Predictive Performance

Both ANN and RF models demonstrated varying predictive performance across the three water quality indicators, reflecting differences in underlying environmental processes and data structure. Overall, the models captured a large proportion of the observed monthly variability, particularly when incorporating temporal dependencies (e.g., lagged predictors) and spatial organization (e.g., subbasin order). Time-series comparisons of observed and predicted values for the independent test dataset (Supplementary Figure 1) show that both models successfully reproduce the general seasonal dynamics across the basin. However, because the results are presented as averages across all monitoring sites, local variability and extreme events are partially smoothed. As a result, while the models capture broad temporal trends, they exhibit limited ability to reproduce peak magnitudes and short-term fluctuations. Model performance was more robust at seasonal scales, suggesting that large-scale hydrological and climatic drivers are better represented than high-frequency variability. Notable differences in predictive accuracy among NTU, NO_3, and PO_4 further highlight the distinct environmental controls governing each water quality parameter.

The RF model demonstrated good predictive performance across the three water quality indicators, with clear differences in model skill among variables (Table 2). Model performance was highest for NTU, achieving an R^2 of 0.75, indicating that a substantial proportion of the observed temporal variability was successfully captured. The corresponding error metrics (RMSE = 41.31; MAE = 27.88) suggest that, although some deviations remain, the model effectively reproduces the overall temporal dynamics of turbidity across the monitoring network. For NO_3, the RF model also showed robust performance, with an R^2 of 0.70 (RMSE = 0.37; MAE = 0.21). The model captured the general monthly variability well, although some discrepancies were observed during periods of higher concentrations, indicating a reduced ability to fully represent peak events. In comparison, predictive performance for PO_4 was lower

but still moderate, with an R^2 of 0.58 (RMSE = 0.06; MAE = 0.03). While the model was able to reproduce general temporal trends, its ability to explain variability was comparatively limited, suggesting that PO_4 dynamics may be influenced by additional processes not fully represented by the available predictors. These results indicate that the RF model performs most effectively for NTU, followed by NO_3, with comparatively weaker but still meaningful performance for PO_4. This pattern highlights differences in the underlying environmental controls of each parameter and suggests that turbidity is more strongly linked to the included predictors, whereas nutrient dynamics—particularly PO_4—are governed by more complex or localized processes.

The ANN models captured the broad monthly and seasonal dynamics of NTU, NO_3, and PO_4 across the monitoring network, although predictive performance varied among indicators (Table 2). For NTU, the ANN demonstrated moderate predictive skill ($R^2 = 0.64$), lower than that of the RF model. The model successfully reproduced the general temporal and seasonal patterns, although some deviations remained in representing peak magnitudes. The corresponding error values (RMSE = 49.20; MAE = 34.54) indicate that, while overall trends were captured, the model still exhibited limitations in accurately reproducing high turbidity events and short-term fluctuations. For NO_3, the ANN showed weaker predictive performance, with an R^2 of 0.21 (RMSE = 0.60; MAE = 0.42). Although the model was able to follow general temporal trends, it exhibited limited ability to capture variability, particularly during periods of elevated concentrations. For PO_4, predictive performance remained relatively low, with an R^2 of 0.17 (RMSE = 0.08; MAE = 0.05). While the error values were small, this largely reflects the low variability and narrow concentration range of PO_4 rather than strong predictive capability. The ANN captured general trends but failed to reproduce meaningful temporal variability. Overall, these results indicate that the ANN can capture dominant temporal patterns, particularly for variables with stronger seasonal signals such as NTU. However, its ability to represent high variability and extreme events remains limited, especially compared to the RF model. This limitation likely reflects both the inherent variability of environmental data and the sensitivity of ANN models to sample size and data structure. Consequently, the ANN performance metrics should be interpreted as reflecting the model's ability to capture general temporal dynamics rather than precise site-level concentrations.

These results highlight that model performance varies not only by modelling approach but also by water-quality parameter, reflecting differences in the underlying environmental controls and the extent to which these processes are captured by the available predictors.

Table 2: Evaluation matrix for Random Forest and Artificial Neural Networks (MAE, RMSE, R^2):

Measures	**Predicted NTU (RF)**	**Predicted NO_3 (RF)**	**Predicted PO_4 (RF)**	**Predicted NTU (ANN)**	**Predicted NO_3 (ANN)**	**Predicted PO_4 (ANN)**
$\boldsymbol{R^2}$	0.75	0.70	0.58	0.64	0.21	0.17
RMSE	41.31	0.37	0.06	49.20	0.60	0.08
MAE	27.88	0.21	0.03	34.54	0.42	0.05

3.3 Land use and hydrological determinants

Both machine learning approaches were used to explore the relative impact of each feature on water quality in each sub-basin using a complementary feature-attribution framework. For both the RF and ANN models, feature importance was quantified using the mean absolute SHAP values to quantify feature importance and Spearman

correlations (with both observed values and out-of-fold predictions) to characterize the direction of association between predictors and water-quality indicators. Together, these complementary approaches enabled the identification of both the magnitude of feature contributions and the direction of association across models.

For the RF model, the features with the greatest explanatory power were identified based on mean absolute SHAP values (Supplementary Table 3). The results reveal pronounced differences in key predictors across the three water-quality indicators. For NTU, forest percent of the sub-basin emerged as the most influential predictor (SHAP = 24.81), followed by precipitation (22.58) and water-flow velocity (11.99), with pasture extent (8.37) and soil erosion (6.13) also contributing substantially. Most of the dominant predictors exhibited positive associations with NTU, particularly precipitation and flow velocity, indicating that hydrological variables are strongly associated with turbidity variability in the model. The positive association observed for forest cover may appear counterintuitive, as forests are typically expected to reduce sediment transport. However, in this study region, forested areas are predominantly located in upstream or high-rainfall zones, where increased precipitation and runoff can elevate turbidity levels. Thus, forest cover likely acts as a proxy for hydro-climatic conditions rather than a direct causal driver. In contrast, variables such as soil erosion, road density, and urban extent exhibited negative or weak directional signals, suggesting more localized or context-dependent influences. For NO_3, precipitation (SHAP = 0.137) was the most influential predictor, followed by road density (0.081) and pasture extent (0.077). Unlike NTU, NO_3 results showed a mixture of positive and negative relationships. Pasture extent displayed a clear negative association with NO_3 (Spearman ≈ –0.45 for out-of-fold (OOF) predictions), while several anthropogenic variables (population density, urban extent, and road density) showed consistent positive relationships. These patterns suggest that nitrate concentrations are influenced by both land-use intensity and hydrological transport processes, with different predictors exerting opposing effects depending on local conditions. For PO_4, overall feature importance values were substantially lower than for NTU and NO_3, indicating weaker model sensitivity to the available predictors. Precipitation remained the dominant variable (SHAP = 0.029), followed by pasture extent and flow velocity (both ≈ 0.013). However, most predictors exhibited weak and often inconsistent directional relationships, with Spearman correlations generally close to zero. This suggests that PO_4 dynamics are less strongly conditioned by the static and semi-static watershed characteristics included in the model, and may instead be influenced by localized processes, point sources, or short-term variability not captured by the available predictors.

Feature-importance from the ANN models was based on the mean absolute SHAP values, Spearman correlations, and directionality indicators (Supplementary Table 4). Given the comparatively lower predictive performance of the ANN model (Section 3.2), these feature-attribution results should be interpreted with caution. The results show both the magnitude and directionality of each feature's contribution to the model outputs. For NTU, the ANN model indicated forest percent of sub-basin (SHAP = 19.335) and urban extent (18.909) as the most influential features, followed by water-flow velocity (14.968) and human footprint in 2009 (14.583). Forest cover shows a negative direction, indicating that higher forest cover is associated with lower predicted turbidity values in the model, whereas urban and agricultural features generally contribute to higher NTU. Human footprint change (1993-2009) also appears among the top predictors, indicating that long-term land-use transitions significantly impact turbidity. The ANN's emphasis on land-cover composition suggest a more nonlinear sensitivity to vegetation and urbanization patterns compared to RF.

The differences observed between the RF and ANN models in terms of predictor directionality, particularly for variables such as forest cover, are not unexpected, as the two approaches capture relationships in fundamentally different ways. The RF model, based on tree-based partitioning, is generally more robust to noise and can effectively represent complex, nonlinear interactions in environmental data. In contrast, ANN models are more sensitive to data scaling, sample size, and parameter configuration, which may lead to different representations of variable effects. Given the relatively limited dataset and the complexity of environmental processes, these differences reflect model-specific representations of the underlying relationships on the underlying system dynamics.

For NO_3, the ANN identifies water-flow velocity (SHAP = 0.224) and Human footprint (2009) (0.197) as the two strongest predictors, followed by pasture extent (0.194) and road density (0.163). In the ANN model, most high-ranking predictors display positive directionality, indicating that higher anthropogenic pressure is associated with higher predicted NO_3 concentrations. Spearman correlations similarly show moderate positive associations for population density and footprint variables, while spatial and urban indicators exhibit mixed or weak relationships. For PO_4, the ANN results exhibit uniformly small SHAP magnitudes, with pasture extent (0.059), population density

(0.032), and soil erosion (0.030) emerging as the highest but still modest contributors. Like the RF model, most predictors show very low explanatory power and weak Spearman correlations, confirming the limited explanatory power of the model for PO_4 and the difficulty of capturing its dynamics using the available predictors.

Discussion

It is important to note that the modelling framework applied in this study is statistical rather than process based. While the models identify consistent associations between environmental predictors and water-quality responses, they do not explicitly represent underlying hydrological or biogeochemical processes such as sediment transport, groundwater contributions, or nutrient cycling. As such, interpretations of feature importance should be understood as indicative of dominant patterns in the data, rather than direct representations of mechanistic pathways.

4.1 Commonalities of feature importance

To compare the influence of sub-basin features on monthly site-level NTU, NO_3, and PO_4 concentrations, SHAP feature-direction analysis for RF and ANN models was performed. Despite fundamental differences in model structure and learning strategy, both ANN and RF identified broadly consistent sets of influential predictors for each water-quality variable, indicating that some predictor–response relationships are consistent across model structures, particularly for NTU, while others show substantial divergence. Agreement in feature ranking between the two approaches varied by constituent. Rank concordance was moderate for turbidity (NTU; $\rho = 0.54$, $p = 0.06$, $n = 12$), weaker for phosphate (PO_4; $\rho = 0.32$, $p = 0.25$), and negligible for nitrate (NO_3; $\rho = 0.01$, $p = 0.95$). These differences reflect increasing complexity and highlight the importance of using multiple modelling approaches to improve confidence in interpretation when working with heterogeneous citizen-science data. Importantly, the Spearman statistics were used as descriptive indicators of rank similarity and do not affect model performance or predictive skill, which were evaluated independently.

A key consideration in interpreting these results is the inherent uncertainty associated with citizen-science water-quality measurements. Previous validation work in comparable East African contexts (Moshi et al., 2023) demonstrates that such data retain strong classification accuracy and inter-observer agreement, particularly for turbidity and nutrient concentration classes. In this context, the modelling framework adopted here is designed to extract dominant patterns and relative driver importance rather than precise point estimates. The consistency of identified predictors across two independent modelling approaches (RF and ANN) further suggests that the models are capturing robust environmental signals rather than artefacts of observational noise.

When examined beyond exact rank positions, predictors identified by both ANN and RF clustered into a small number of consistent functional categories across all three constituents. This included catchment buffering capacity (e.g., forest cover and soil erosion indicators), diffuse agricultural sources (pasture and cropland extent), aggregated human pressure (human footprint indices, population metrics, road density, and night-time lights), and local river hydrology represented by precipitation and citizen-science estimates of water-flow velocity.

It is important to note that the feature importance and SHAP analyses presented here explore statistical associations between environmental variables and model predictions, and not direct causal mechanisms. Given the complex interactions among land use, hydrology, and biogeochemical processes in river systems, these relationships should be interpreted as indicative of patterns within the data rather than direct evidence of process-level causation.

For NTU, the major positive drivers were primarily associated with land-use variables, with pasture extent showing the strongest positive SHAP contribution (mean SHAP value approaching 1.2102, absolute mean SHAP value = 8.366) (Table 3 & Supplementary Table 3), indicating that pasture extent is strongly positively associated with turbidity in the model. By contrast, forest cover (-0.8768), historical human footprint indices (-0.4845), and soil-erosion intensity exhibited pronounced negative SHAP values (approximately −0.3403), suggesting an association with lower sediment inputs to the river and its tributaries.

In comparison, NO_3 was predominantly influenced by demographic and urbanization-related drivers, including population density, urban extent, and contemporary human footprint, suggesting that NO_3 concentrations are more sensitive to domestic wastewater discharge, urban non-point source inputs, and population aggregation. Negative contributions for NO_3 were largely associated with several land-cover variables. The positive and negative impact

intensities of each feature are broadly comparable between modelling approaches. PO_4 exhibited a weak and somewhat variable response pattern, although these relationships should be interpreted with caution given the limited model performance: precipitation and pasture extent had strong negative effects (Table 3), whereas road density, population density, population counts and human-footprint variables consistently contributed positively.

Differences in rank concordance among water quality indicators reflect constituent-specific sensitivities to these functional drivers. NTU was most strongly associated with hydrological and landscape variables, including forest cover, precipitation, and flow velocity, leading to greater agreement between models and more spatially coherent responses. In contrast, NO_3 exhibited minimal rank concordance, consistent with its regulation by multiple diffuse sources, legacy and contemporary human pressures, and complex nonlinear interactions that vary across space. Phosphate showed intermediate behaviour, reflecting its dual association with sediment-linked transport and localized anthropogenic inputs. Together, these patterns suggest that some sub-basins are structurally predisposed to event-driven responses, whereas others experience persistent background loading modulated by hydrological conditions.

Table 3: Mean SHAP values and impact directions of sub-basin characteristics on NTU, NO_3, and PO_4 in Random Forest (RF) and Artificial Neural Network (ANN) models (mean SHAP values were used to interpret whether predictors increase or decrease water-quality variables)

NTU			
RF		**ANN**	
Feature	**Mean SHAP**	**Feature**	**Mean SHAP**
Pasture extent of sub-basin	1.2102	Human footprint of sub-basin 2009	4.3518
cropland extent of sub-basin	0.4224	Human footprints change of sub-basin 1993-2009	2.2434
precipitation	0.3541	Pasture extent of sub-basin	1.3424
Human footprints change of sub-basin 1993-2009	0.2238	precipitation	1.2043
Night-time lights of sub-basin	0.1484	citizen scientist estimate of water flow velocity	0.4167
Population density of sub-basin	-0.0117	Population counts of sub-basin	0.1798
Urban extent of sub-basin	-0.1675	Night-time lights of sub-basin	-0.0623
Road density of sub-basin	-0.1912	cropland extent of sub-basin	-0.1419
Population counts of sub-basin	-0.2355	Population density of sub-basin	-0.5552
Soil Erosion of sub-basin	-0.3403	Urban extent of sub-basin	-0.9812
Human footprint of sub-basin 2009	-0.4845	Road density of sub-basin	-1.4691
citizen scientist estimate of water flow velocity	-0.5110	Soil Erosion of sub-basin	-1.6318
forest percent of sub-basin	-0.8768	forest percent of sub-basin	-1.8100
NO3			
RF		**ANN**	
Feature	**Mean SHAP**	**Feature**	**Mean SHAP**
forest percent of sub-basin	0.0054	Road density of sub-basin	0.0772
Human footprint of sub-basin 2009	0.0047	Pasture extent of sub-basin	0.0456
citizen scientist estimate of water flow velocity	0.0039	Soil Erosion of sub-basin	0.0282
precipitation	0.0034	Urban extent of sub-basin	0.0239
Human footprints change of sub-basin 1993-2009	0.0011	citizen scientist estimate of water flow velocity	0.0214
Road density of sub-basin	0.0002	precipitation	0.0152

Pasture extent of sub-basin	0.0000	Night-time lights of sub-basin	0.0113
cropland extent of sub-basin	-0.0006	cropland extent of sub-basin	0.0109
Night-time lights of sub-basin	-0.0008	Population counts of sub-basin	0.0017
Soil Erosion of sub-basin	-0.0012	Human footprints change of sub-basin 1993-2009	-0.0044
Population counts of sub-basin	-0.0036	forest percent of sub-basin	-0.0172
Population density of sub-basin	-0.0059	Population density of sub-basin	-0.0517
Urban extent of sub-basin	-0.0065	Human footprint of sub-basin 2009	-0.0558
PO4			
RF		**ANN**	
Feature	**Mean SHAP**	**Feature**	**Mean SHAP**
forest percent of sub-basin	0.0008	Human footprint of sub-basin 2009	0.0090
Population counts of sub-basin	0.0005	Population density of sub-basin	0.0072
citizen scientist estimate of water flow velocity	0.0005	Human footprints change of sub-basin 1993-2009	0.0044
Urban extent of sub-basin	0.0004	Population counts of sub-basin	0.0034
Soil Erosion of sub-basin	0.0004	Road density of sub-basin	0.0034
Human footprint of sub-basin 2009	0.0004	Pasture extent of sub-basin	0.0031
Human footprints change of sub-basin 1993-2009	0.0003	cropland extent of sub-basin	0.0026
Population density of sub-basin	0.0003	Night-time lights of sub-basin	0.0001
Road density of sub-basin	0.0000	citizen scientist estimate of water flow velocity	-0.0005
cropland extent of sub-basin	-0.0002	precipitation	-0.0014
Night-time lights of sub-basin	-0.0003	Soil Erosion of sub-basin	-0.0053
Pasture extent of sub-basin	-0.0012	forest percent of sub-basin	-0.0059
precipitation	-0.0030	Urban extent of sub-basin	-0.0107

4.2 Sub-basin sensitivity to climate variables

Using site-level SHAP values and correlation analyses, it was possible to compare the sensitivity of different subbasins to changes in precipitation, given that precipitation in East Africa is expected to increase in intensity and duration (Ayugi et al., 2021; Wang et al., 2025). Both RF and ANN models consistently indicate that the direction and magnitude of precipitation associations with water quality are strongly site dependent, reflecting differences in geomorphology, land use, and hydrological connectivity along the river continuum. The models identify distinct upstream, transitional, and downstream patterns, underscoring the importance of local controls in shaping turbidity dynamics. For example, land use and vegetation cover influence how precipitation translates into runoff, sediment transport, and nutrient export. However, the limited availability of time-varying predictors remains a constraint, and future work should incorporate additional dynamic variables to better capture temporal processes.

For NTU, the upper and parts of the middle basin, including headwater and transitional sites along the Nyangores and Amala rivers, precipitation is predominantly associated with positive SHAP values, indicating that increased rainfall is associated with higher predicted NTU values. This pattern is observed at sites such as at Lower mara, Amala River upper bridge, Amala River at Mulot bridge D/S, Nyangores River at Bomet Town and Mto manchira makutano romakendo (Table 4). This response suggests the high sensitivity of upstream reaches to precipitation-related changes in turbidity, consistent with conditions where intense rainfall can enhance soil erosion, hillslope runoff, and rapid sediment mobilization; a matter that is exacerbated by unterraced hillslope farming and murran/earth roads network.

In contrast, sites located in the downstream basin predominantly exhibit a negative relationship between precipitation and NTU, characterized by negative SHAP values under increasing precipitation. This pattern may be associated with the increased channel width and discharge capacity downstream, which are consistent with dilution effects, as well as

stronger hydraulic connectivity with floodplains that can promote sediment deposition during high-flow events, thereby limiting the transport of fine sediments to monitoring sites. In addition, longer flow paths and gentler channel slopes in the downstream basin further facilitate sediment settling before reaching observation points.

Differently, precipitation was associated with a negative SHAP values for NO_3 for headwater sites, with an increasing SHAP values in downstream sites. Sites such as Mto manchira makutano romakendo, Nyangores River at Chemomul Bridge, and Amala River upper bridge, have negative SHAP values, indicating that storm events are associated with lower predicted concentrations, consistent with dilution effects. This pattern suggests hydrological flushing in steep catchments with limited nutrient storage and short flow paths (Table 5). In mid- and lower-basin sites, including Mto Mara Kirumi Darajani, Mto Somoche Nyiboko, and Mto Mori keisangora, generally with increased agricultural activities, rainfall is associated with increased NO_3 concentrations, consistent with mobilization of fertilizer residues, soil nitrogen, and potential shallow groundwater inputs. The stronger connectivity between hillslopes, floodplains, and channel systems can facilitate the storm-driven transfer of nitrate, leading to concentration increases during rainfall events.

In contrast to NTU and NO_3, both modelling approaches demonstrated limited explanatory power for PO_4, indicating that the identified relationships should be interpreted with caution. The influence of precipitation on PO_4 responses is relatively balanced across the basin, with no clear or systematic distinction between upstream and downstream reaches. Although precipitation is associated with positive SHAP values at several upstream sites, indicating that rainfall s associated with increased phosphate concentrations, consistent with rain-driven inputs, both positive and negative SHAP values of comparable magnitude are also observed at downstream sites. For example, at upstream locations such as the Kisangura Darajani, Lower Mara, Amala River downstream of Mulot Bridge, and the Isei river bridge D/S (Table 6), rainfall-induced increases in phosphate (PO_4) concentrations may reflect the mobilization of phosphorus from wastewater discharges, fertilizer residues and near-channel sources in populated areas, (Sharpley, 1995; Withers & Jarvie, 2008), although this interpretation remains uncertain given the low model performance and weak feature importance values.

Table 4: Site-specific precipitation sensitivity of NTU inferred from mean SHAP values in Random Forest (RF) and Artificial Neural Network models (ANN) (from upper sub-basins to lower sub-basins, from Figure 1, Supplementary Table 2).

NTU			
RF		**ANN**	
Site Name	**Mean SHAP precip**	**Site Name**	**Mean SHAP precip**
Bwawa La Manchira, Intake	-1.542	Bwawa La Manchira, Intake	-5.782
Kisangura Darajani	0.958	Kisangura Darajani	5.056
Mto manchira makutano romakendo	2.618	Mto manchira makutano romakendo	5.717
Upper Amala at Timoi	-1.434	Upper Amala at Timoi	-2.829
Isei river bridge D/S	-1.017	Isei river bridge D/S	4.415
Amala river upper bridge	2.882	Amala river upper bridge	6.688
Amala River at Mulot bridge D/S	1.904	Amala River at Mulot bridge D/S	6.454
Upper Nyangores at TMC bridge	-3.285	Upper Nyangores at TMC bridge	-4.681
Upper Nyangores at Lelbanga bridge	0.229	Upper Nyangores at Lelbanga bridge	2.517

Nyangores River at Chemomul Bridge	1.042	Nyangores River at Chemomul Bridge	6.021
Nyangores River at Bomet Town	2.404	Nyangores River at Bomet Town	2.317
Mto Rubana Darajani Makundusi	-7.186	Mto Rubana Darajani Makundusi	-10.421
Illdungushu siana Wrua @ Nalengitomon	4.223	Illdungushu siana Wrua @ Nalengitomon	10.724
Talek river talek gate	-1.481	Talek river talek gate	-0.642
Lower mara	5.68	Lower mara	14.384
Marasienta - Mara river	1.494	Marasienta - Mara river	0.889
Mto Nyamonyo Mlimani Wegero	-2.356	Mto Nyamonyo Mlimani Wegero	2.276
Mto Mara Nyamerambaro Kiriba	-2.526	Mto Mara Nyamerambaro Kiriba	-4.775
Mto Nyarusondobiro Kwa Juma Magabe	-0.176	Mto Nyarusondobiro Kwa Juma Magabe	0.527
Mto Mara Kitasakwa Kwiswa	0.261	Mto Mara Kitasakwa Kwiswa	-3.716
Mto Nyamakendo Darajani	-0.566	Mto Nyamakendo Darajani	1.689
Mto tobora koreri	-0.706	Mto tobora koreri	1.267
Mto Mara Kivuko	-0.225	Mto Mara Kivuko	0.066
Mto Tobora Makutano	2.269	Mto Tobora Makutano	3.442
Mto Tigite Juu	2.732	Mto Tigite Juu	5.036
Mto Tigite Chini Kewanja Darajani	-1.361	Mto Tigite Chini Kewanja Darajani	1.138
Remaiso(North Mara Before Discharge)	-0.891	Remaiso(North Mara Before Discharge)	2.806
Mto Nyaruherere Darajan Musati	-7.035	Mto Nyaruherere Darajan Musati	-7.972
Mto Somoche Darajani Kenyana	-1.571	Mto Somoche Darajani Kenyana	-1.514
Nyantare(North Mara After Dischanging)	-2.787	Nyantare(North Mara After Dischanging)	-4.285
Mto Somoche Nyiboko	1.057	Mto Somoche Nyiboko	-0.645
Tigite Rechawi Matongo	-5.757	Tigite Rechawi Matongo	-3.473
Mto Nyamitita Darajani	-0.587	Mto Nyamitita Darajani	3.516
Mto Somoche Nyamatare	-5.882	Mto Somoche Nyamatare	-13.786
Mto Mara, Bisarwi Burongo	1.185	Mto Mara, Bisarwi Burongo	6.779
Mto Mara Kirumi Darajani	0.279	Mto Mara Kirumi Darajani	2.844
Mto Mori -keisangora	-1.379	Mto Mori -keisangora	1.247
Mto Mori Buganjo/Gamasara	0.821	Mto Mori Buganjo/Gamasara	2.985

Mto Mori, Utegi Darajani	-1.792	Mto Mori, Utegi Darajani	-3.32
Mto Mori Wamaya-Kirogo	-1.456	Mto Mori Wamaya-Kirogo	-3.552

Table 5: Site-specific precipitation sensitivity of NO_3 inferred from mean SHAP values in Random Forest (RF) and Artificial Neural Network models (ANN) (from upper sub-basins to lower sub-basins, from Figure 1, Supplementary Table 2).

NO_3			
RF		**ANN**	
Site Name	**Mean SHAP precip**	**Site Name**	**Mean SHAP precip**
Bwawa La Manchira, Intake	-0.011	Bwawa La Manchira, Intake	0.05
Kisangura Darajani	0.012	Kisangura Darajani	-0.034
Mto manchira makutano romakendo	-0.006	Mto manchira makutano romakendo	-0.02
Upper Amala at Timoi	-0.011	Upper Amala at Timoi	0.043
Isei river bridge D/S	-0.01	Isei river bridge D/S	0
Amala river upper bridge	-0.005	Amala river upper bridge	-0.021
Amala River at Mulot bridge D/S	-0.003	Amala River at Mulot bridge D/S	-0.022
Upper Nyangores at TMC bridge	-0.01	Upper Nyangores at TMC bridge	0.04
Upper Nyangores at Lelbanga bridge	-0.013	Upper Nyangores at Lelbanga bridge	0.026
Nyangores River at Chemomul Bridge	-0.003	Nyangores River at Chemomul Bridge	-0.006
Nyangores River at Bomet Town	0.006	Nyangores River at Bomet Town	0.117
Mto Rubana Darajani Makundusi	-0.027	Mto Rubana Darajani Makundusi	0.076
Illdungushu siana Wrua @ Nalengitomon	-0.01	Illdungushu siana Wrua @ Nalengitomon	-0.039
Talek river talek gate	-0.004	Talek river talek gate	-0.019
Lower mara	0	Lower mara	-0.104
Marasienta - Mara river	-0.005	Marasienta - Mara river	-0.015
Mto Nyamonyo Mlimani Wegero	-0.004	Mto Nyamonyo Mlimani Wegero	0.024
Mto Mara Nyamerambaro Kiriba	-0.017	Mto Mara Nyamerambaro Kiriba	-0.056
Mto Nyarusondobiro Kwa Juma Magabe	0	Mto Nyarusondobiro Kwa Juma Magabe	0.033
Mto Mara Kitasakwa Kwiswa	0.035	Mto Mara Kitasakwa Kwiswa	0.038
Mto Nyamakendo Darajani	0.013	Mto Nyamakendo Darajani	0.002
Mto tobora koreri	0.001	Mto tobora koreri	0.005

Mto Mara Kivuko	-0.004	Mto Mara Kivuko	0.004
Mto Tobora Makutano	-0.011	Mto Tobora Makutano	-0.01
Mto Tigite Juu	-0.004	Mto Tigite Juu	-0.006
Mto Tigite Chini Kewanja Darajani	-0.002	Mto Tigite Chini Kewanja Darajani	0.013
Remaiso(North Mara Before Discharge)	0.004	Remaiso(North Mara Before Discharge)	0.036
Mto Nyaruherere Darajan Musati	-0.007	Mto Nyaruherere Darajan Musati	0.059
Mto Somoche Darajani Kenyana	0.001	Mto Somoche Darajani Kenyana	0.051
Nyantare(North Mara After Dischanging)	-0.007	Nyantare(North Mara After Dischanging)	0.067
Mto Somoche Nyiboko	0.051	Mto Somoche Nyiboko	0.054
Tigite Rechawi Matongo	-0.01	Tigite Rechawi Matongo	0.027
Mto Nyamitita Darajani	-0.016	Mto Nyamitita Darajani	-0.024
Mto Somoche Nyamatare	0	Mto Somoche Nyamatare	0.076
Mto Mara, Bisarwi Burongo	0.006	Mto Mara, Bisarwi Burongo	-0.041
Mto Mara Kirumi Darajani	0.002	Mto Mara Kirumi Darajani	0.026
Mto Mori -keisangora	0.036	Mto Mori -keisangora	0.064
Mto Mori Buganjo/Gamasara	-0.001	Mto Mori Buganjo/Gamasara	0.001
Mto Mori, Utegi Darajani	-0.004	Mto Mori, Utegi Darajani	0.016
Mto Mori Wamaya-Kirogo	-0.008	Mto Mori Wamaya-Kirogo	0.028

Table 6: Site-specific precipitation sensitivity of PO_4 inferred from mean SHAP values in Random Forest (RF) and Artificial Neural Network models (ANN) (from upper sub-basins to lower sub-basins, from Figure 1, Supplementary Table 2).

PO_4			
RF		**ANN**	
Site Name	**Mean SHAP precip**	**Site Name**	**Mean SHAP precip**
Bwawa La Manchira, Intake	-0.0034	Bwawa La Manchira, Intake	0.001
Kisangura Darajani	0.0005	Kisangura Darajani	0.003
Mto manchira makutano romakendo	-0.0003	Mto manchira makutano romakendo	-0.007
Upper Amala at Timoi	-0.0033	Upper Amala at Timoi	-0.005
Isei river bridge D/S	0.001	Isei river bridge D/S	0.005
Amala river upper bridge	0.0014	Amala river upper bridge	0.003

Amala River at Mulot bridge D/S	0.0022	Amala River at Mulot bridge D/S	0.002
Upper Nyangores at TMC bridge	-0.0033	Upper Nyangores at TMC bridge	-0.012
Upper Nyangores at Lelbanga bridge	-0.0017	Upper Nyangores at Lelbanga bridge	-0.007
Nyangores River at Chemomul Bridge	0.0012	Nyangores River at Chemomul Bridge	0.003
Nyangores River at Bomet Town	-0.0009	Nyangores River at Bomet Town	0.005
Mto Rubana Darajani Makundusi	-0.0062	Mto Rubana Darajani Makundusi	-0.001
Illdungushu siana Wrua @ Nalengitomon	0.0072	Illdungushu siana Wrua @ Nalengitomon	0.008
Talek river talek gate	-0.005	Talek river talek gate	-0.003
Lower mara	0.0078	Lower mara	0.009
Marasienta - Mara river	-0.0042	Marasienta - Mara river	-0.001
Mto Nyamonyo Mlimani Wegero	0.0001	Mto Nyamonyo Mlimani Wegero	0.005
Mto Mara Nyamerambaro Kiriba	-0.0018	Mto Mara Nyamerambaro Kiriba	-0.007
Mto Nyarusondobiro Kwa Juma Magabe	-0.0009	Mto Nyarusondobiro Kwa Juma Magabe	-0.013
Mto Mara Kitasakwa Kwiswa	0	Mto Mara Kitasakwa Kwiswa	0.004
Mto Nyamakendo Darajani	0.0001	Mto Nyamakendo Darajani	-0.003
Mto tobora koreri	-0.0009	Mto tobora koreri	-0.003
Mto Mara Kivuko	0	Mto Mara Kivuko	0.002
Mto Tobora Makutano	0.0004	Mto Tobora Makutano	0.008
Mto Tigite Juu	-0.0002	Mto Tigite Juu	0.003
Mto Tigite Chini Kewanja Darajani	-0.0001	Mto Tigite Chini Kewanja Darajani	0.001
Remaiso(North Mara Before Discharge)	0.0008	Remaiso(North Mara Before Discharge)	0.001
Mto Nyaruherere Darajan Musati	-0.0038	Mto Nyaruherere Darajan Musati	-0.003
Mto Somoche Darajani Kenyana	-0.0015	Mto Somoche Darajani Kenyana	0.003
Nyantare(North Mara After Dischanging)	-0.0029	Nyantare(North Mara After Dischanging)	-0.009
Mto Somoche Nyiboko	0.0051	Mto Somoche Nyiboko	0.023
Tigite Rechawi Matongo	0.0039	Tigite Rechawi Matongo	0.004
Mto Nyamitita Darajani	0.0009	Mto Nyamitita Darajani	0.001
Mto Somoche Nyamatare	-0.0078	Mto Somoche Nyamatare	-0.002
Mto Mara, Bisarwi Burongo	-0.0012	Mto Mara, Bisarwi Burongo	0.002

Mto Mara Kirumi Darajani	0.0003	Mto Mara Kirumi Darajani	0.002
Mto Mori -keisangora	0.0001	Mto Mori -keisangora	0.007
Mto Mori Buganjo/Gamasara	0.0003	Mto Mori Buganjo/Gamasara	0
Mto Mori, Utegi Darajani	-0.0015	Mto Mori, Utegi Darajani	0.003
Mto Mori Wamaya-Kirogo	-0.0008	Mto Mori Wamaya-Kirogo	0.01

4. Conclusions

The two machine learning approaches exhibited notable differences in predictive performance, feature sensitivity, and their ability to capture the dominant predictor–response relationships underlying water-quality variations. These differences reflect the distinct structural designs and learning principles of the two models. In terms of overall predictive accuracy, the RF model demonstrated strong performance for NTU and NO_3, and moderate performance for PO_4, achieving R^2 values of 0.75 for NTU, 0.70 for NO_3, and 0.58 for PO_4 (Table 2). The corresponding error metrics further support this performance, with RMSE values of 41.31, 0.37, and 0.06, and MAE values of 27.88, 0.21, and 0.03 for NTU, NO_3, and PO_4, respectively. These results indicate that the RF model captures both temporal variability and complex nonlinear relationships in the data. This strong performance can be attributed to the ensemble learning framework, which reduces variance and enables robust representation of interactions among environmental predictors and citizen-science-derived observations. We note, however, that model performance remains partly influenced by dataset size and structure, which may inherently favor tree-based methods in this context. In addition, the inclusion of temporal features (e.g., lagged and rolling variables) further enhances the model's ability to represent short-term dynamics. In contrast, the ANN exhibited more variable and generally weaker predictive performance across the three indicators. While the ANN achieved moderate performance for NTU ($R^2 = 0.64$), it remained lower than that of the RF model, with higher prediction errors (RMSE = 49.20; MAE = 34.54), indicating reduced accuracy in capturing high-magnitude variability. For NO_3 and PO_4, predictive performance was considerably lower ($R^2 = 0.21$ and 0.17, respectively), with larger associated errors (RMSE = 0.60 and 0.08; MAE = 0.42 and 0.05). These results suggest that the ANN is less effective in capturing variability, particularly for variables influenced by complex spatial heterogeneity or weaker temporal signals. Tree-based ensemble methods such as RF provide more stable and reliable predictive performance in data-limited and spatially structured environmental systems. In contrast, ANN models, despite their flexibility, tend to exhibit higher uncertainty and reduced robustness under such conditions, particularly when modelling variables with strong variability or limited explanatory predictors.

The two approaches showed some similarities in feature importance, likely reflecting their distinct approaches to interpreting data structure. RF emphasizes hydrological and basin-scale variables, such as forest cover, precipitation, and flow velocity, as tree-based models can identify environmental features that exhibit clear threshold effects within the data. This tendency aligns with RF's ability to reconstruct overall temporal trends in all three water-quality indicators. By contrast, ANN is robust for high-dimensional nonlinear transformations, making it more sensitive to socioeconomic variables such as population density, human footprint indices, and road density. However, because ANN shows relatively poor predictive performance overall, its feature-sensitivity and results should be interpreted with caution; although ANN identifies several human-related variables as influential, the model's low R^2 values suggest that its explanatory power may be limited. Despite these differences, the two approaches exhibit some complementarity. RF was better suited for capturing water-quality patterns associated with natural environmental gradients, whereas ANN may detect more complex and nonlinearly interacting socio-ecological influences. Considering both approaches therefore provides a more comprehensive understanding of water-quality dynamics in large, complex transboundary river basins.

This study demonstrates how citizen-science data, when combined with interpretable machine-learning approaches, can be used to identify spatial drivers and rainfall-sensitive water-quality responses in data-scarce river basins. Both ANN and RF results consistently highlight precipitation, flow velocity, and erosion-related features as key predictors of short-term fluctuations, indicating consistent associations between rainfall variability and observed sediment and nutrient patterns. Under conditions of high rainfall variability and more frequent extreme precipitation events, our findings suggest that machine-learning models can effectively capture these event-driven patterns in water-quality

variability and provide indicative information to support water-resource management. By integrating citizen-science observations with large-scale environmental datasets, this study provides a transferable and scalable analytical framework for extracting decision-relevant insights from citizen-science datasets and offers opportunities for pollution detections and targeted intervention in data scarce regions. This provides a pathway for addressing rainfall-induced water-quality risks and supports the development of adaptive monitoring and management strategies in regions where hydrological systems are highly sensitive to climatic change. However, the limited explanatory performance of the models for PO_4 highlights the need for caution in interpreting these relationships and suggests that additional process-relevant variables or hybrid modelling approaches may be required to better capture phosphorus dynamics.

References

Ayugi B, Dike V, Ngoma H, Babaousmail H, Mumo R, Ongoma V (2021) Future changes in precipitation extremes over East Africa based on CMIP6 models. Water 13(17):2358. https://doi.org/10.3390/w13172358

Bargam B, Boudhar A, Kinnard C, Bouamri H, Nifa K, Chehbouni A (2024) Evaluation of the support vector regression (SVR) and the random forest (RF) models accuracy for streamflow prediction under a data-scarce basin in Morocco. Discover Appl Sci 6(6):306. https://doi.org/10.1007/s42452-024-05994-z

Bewket W, Sterk G (2005) Dynamics in land cover and its effect on stream flow in the Chemoga watershed, Blue Nile basin, Ethiopia. Hydrol Process 19(2):445–458. https://doi.org/10.1002/hyp.5542

Bishop I, Boldrini A, Clymans W, Hall C, Moorhouse H, Parkinson S, Scott-Somme K, Thornhill I, Loiselle S (2025) FreshWater Watch: Investigating the health of freshwater ecosystems, from the bottom up. Citizen Sci Theory Pract 10(1):1–15. https : //doi.org/10.5334/cstp.754

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Buytaert W, Zulkafli Z, Grainger S, Acosta L, Alemie TC, Bastiaensen J, De Bièvre B, Bhusal J, Clark J, Dewulf A, Foggin M (2014) Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development. Front Earth Sci 2:26. https://doi.org/10.3389/feart.2014.00026

Dunne T (1979) Sediment yield and land use in tropical catchments. J Hydrol 42(3–4):281–300. https://doi.org/10.1016/0022-1694(79)90052-0

Gao Z, Wang G, Zhu Y, Chen J, Fang L, Ren S, Li J, Wang W, Wang Q (2025) Prediction of water quality parameters and pollution exceedance analysis in typical rivers of semi-arid regions based on interpretable deep learning models. Environ Pollut 126801. https://doi.org/10.1016/j.envpol.2025.126801

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge. https://synapse.koreamed.org/pdf/10.4258/hir.2016.22.4.351

Haykin S (2009) Neural networks and learning machines, 3rd edn. Pearson Education India, New Delhi. https://books.google.com/books?hl=en&lr=&id=ivK0DwAAQBAJ&oi=fnd&pg=PR1&dq=Haykin+S+(2009)+Neural+networks+and+learning+machines,+3rd+edn.+Pearson+Education+India,+New+Delhi&ots=8QQ6IbW8k_&sig=8HVXRgnCjRBfCi9bR9eiI7Ku-Vw

Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, Nicolas J, Peubey C, Radu R, Rozum I, Schepers D, Simmons A, Soci C, Dee D, Thépaut J-N (2023) ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store. https://doi.org/10.24381/cds.f17050d7

John KK, Aloyce WM, Joel KN (2018) Hydrology and hydraulics of the Lubigi wetland in Uganda. Tanzania J Eng Technol 37(1):1–18. https://www.researchgate.net/profile/Aw-Mayo/publication/328942167_Hydrology_and_Hydraulics_of_Lubigi_Wetland_in_Uganda/links/5bec587292851c6b27be2550/Hydrology-and-Hydraulics-of-Lubigi-Wetland-in-Uganda.pdf

Kihwele, E. S. (2021). Water and wildlife in the Serengeti-Mara ecosystem. Doctoral dissertation, University of Groningen. https://doi.org/10.33612/diss.164324240

Komba RS (2022) An assessment of spatial variation in surface water dependence among selected savanna ungulates during the wet season in Serengeti National Park. Doctoral dissertation, The Open University of Tanzania. http://repository.out.ac.tz/id/eprint/3813

Li X, Wang Y, Basu S, Kumbier K, Yu B (2019) A debiased MDI feature importance measure for random forests. arXiv preprint arXiv:1906.10845. https://proceedings.neurips.cc/paper_files/paper/2019/hash/702cafa3bb4c9c86e4a3b6834b45aedd-Abstract.html

Linke S, Lehner B, Ouellet Dallaire C, Ariwi J, Grill G, Anand M, Beames P, Burchard-Levine V, Maxwell S, Moidu H, Tan F, Thieme M (2019) Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Sci Data 6:283. https://doi.org/10.1038/s41597-019-0300-6

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. https://doi.org/10.1145/3233231

Lokman A, Ismail WZW, Aziz NAA (2025) A review of water quality forecasting and classification using machine learning models and statistical analysis. Water 17(15):2243. https://doi.org/10.3390/w17152243

Loiselle S, Bishop I, Moorhouse H, Pilat C, Koelman E, Nelson R, Clymans W, Pratt J, Lewis V (2024) Citizen scientists filling knowledge gaps of phosphate pollution dynamics in rural areas. Environ Monit Assess 196(2):220. https://doi.org/10.1007/s10661-024-12389-5

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mango LM, Melesse AM, McClain ME, Gann D, Setegn S (2011a) Land use and climate change impacts on the hydrology of the upper Mara River Basin, Kenya. Hydrol Earth Syst Sci 15(7):2245–2258. https://doi.org/10.5194/hess-15-2245-2011

Mango LM, Melesse AM, McClain ME, Gann D, Setegn SG (2011b) Hydrometeorology and water budget of the Mara River Basin under land use change scenarios. In: Melesse AM (ed) Nile River Basin: Hydrology, Climate and Water Use. Springer, Dordrecht, pp 39–68. https://doi.org/10.1007/978-94-007-0689-7_2

McClain ME, Subalusky AL, Anderson EP, Dessu SB, Melesse AM, Ndomba PM, Mtamba JO, Tamatamah RA, Mligo C (2014) Comparing flow regime, channel hydraulics, and biological communities to infer flow–ecology relationships in the Mara River of Kenya and Tanzania. Hydrol Sci J 59(3–4):725–747. https://doi.org/10.1080/02626667.2013.853121

Meybeck M (2003) Global analysis of river systems: from Earth system controls to Anthropocene syndromes. Philos Trans R Soc B Biol Sci 358(1440):1935–1955. https://doi.org/10.1098/rstb.2003.1379

Moshi HA, Shilla DA, Brehim J, Kimirei I, O'Reilly C, Loiselle S (2023) Sustainable management of the african great lake coastal areas: motivations and perspectives of community citizen scientists. Environ Manage 72(3):473–487. https://doi.org/10.1007/s00267-023-01824-x

Niang, I., Ruppel, O. C., Abdrabo, M. A., Essel, C., Lennard, C., Padgham, J., Urquhart, P., & Descheemaeker, K. K. E. (2014). Chapter 22: Africa. In V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, & K. J. Mach (Eds.), *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment* (pp. 1199-1265). (Intergovernmental Panel on Climate Change). https://doi.org/10.1017/CBO9781107415386.002

Nishat MH, Khan MHRB, Ahmed T, Hossain SN, Ahsan A, El-Sergany MM, Shafiquzzaman M, Imteaz MA, Alresheedi MT (2025) Comparative analysis of machine learning models for predicting water quality index in Dhaka's rivers of Bangladesh. Environ Sci Eur 37(1):31. https://doi.org/10.1186/s12302-025-01078-w

Nokelainen O, Lauha P, Andrejeff S, Hänninen J, Inkinen J, Kallio A, Lehto HJ, Mutanen M, Paavola R, Schiestl-Aalto P, Somervuo P (2024) A mobile application–based citizen science product to compile bird observations. Citizen Sci Theory Pract 9(1):1–12. https://doi.org/10.5334/cstp.710.

Nyairo WN, Owuor PO, Kengara FO (2015) Effect of anthropogenic activities on the water quality of Amala and Nyangores tributaries of River Mara in Kenya. Environ Monit Assess 187(11):691. https://doi.org/10.1007/s10661-015-4913-8

Nyamweya CS, Natugonza V, Taabu-Munyaho A, Aura CM, Njiru JM, Ongore C, Mangeni-Sande R, Kashindye BB, Odoli CO, Ogari Z, Kayanda R (2020) A century of drastic change: Human-induced changes of Lake Victoria fisheries and ecology. Fish Res 230:105564. https://doi.org/10.1016/j.fishres.2020.105564

Odero JO (2024) Land use effects on spatial and seasonal variations of water quality in Thika River Sub-Catchment, Upper Tana Basin, Kenya. Doctoral dissertation, Kenyatta University. https://ir-library.ku.ac.ke/bitstreams/7ca98e0c-7919-43cb-b418-ac0260a46cf9/download

Odenyo VAO (2018) Environmental impacts of population and cultivated ecosystems on the water resources of the Lake Victoria basin: a review in relation to water resources planning and management. J Environ Sustain Adv Res 4:1–12. https://journals.chuka.ac.ke/index.php/jesar/article/view/68

Palmer MA, Reidy Liermann CA, Nilsson C, Flörke M, Alcamo J, Lake PS, Bond N (2008) Climate change and the world's river basins: anticipating management options. Front Ecol Environ 6(2):81–89. https://doi.org/10.1890/060148

Pan D, Deng Y, Yang SX, Gharabaghi B (2025) Recent advances in remote sensing and artificial intelligence for river water quality forecasting: a review. Environments 12(5):158. https://doi.org/10.3390/environments12050158

Parmar KS, Bhardwaj R (2015) Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management. Environ Sci Pollut Res 22(1):397–414. https://doi.org/10.1007/s11356-014-3346-1

Rowell DP, Booth BB, Nicholson SE, Good P (2015) Reconciling past and future rainfall trends over East Africa. J Clim 28(24):9768–9788. https://doi.org/10.1175/JCLI-D-15-0140.1

Sharpley A (1995) Identifying sites vulnerable to phosphorus loss in agricultural runoff. J Environ Qual 24(5):947–951. https://doi.org/10.2134/jeq1995.00472425002400050024x

Soetan O, Nie J, Polius K, Feng H (2024) Application of time series and multivariate statistical models for water quality assessment and pollution source apportionment in an Urban River, New Jersey, USA. Environ Sci Pollut Res 31(52):61643–61659. https://doi.org/10.1007/s11356-024-35330-2

Taylor J, Graham M, Louw A, Lepheana A, Madikizela B, Dickens C, Chapman DV, Warner S (2022) Social change innovations, citizen science, miniSASS and the SDGs. Water Policy 24(5):708–717. https://doi.org/10.2166/wp.2021.264

UNEP (2016) A snapshot of the world's water quality: towards a global assessment. United Nations Environment Programme, Nairobi. https://www.wwqa-documentation.info/assets/policy_brief_unep_wwqa_web3.pdf

UNEP (2024) Progress on ambient water quality: mid-term status of SDG Indicator 6.3.2 and acceleration needs, with a special focus on health. United Nations Environment Programme, Nairobi. https://www.unwater.org/sites/default/files/2024-08/SDG6_Indicator_Report_632_Progress-on-Ambient-Water-Quality_2024_EN_0.pdf

Vörösmarty CJ, McIntyre PB, Gessner MO, Dudgeon D, Prusevich A, Green P, Glidden S, Bunn SE, Sullivan CA, Liermann CR, Davies P (2010) Global threats to human water security and river biodiversity. Nature 467(7315):555–561. https://doi.org/10.1038/nature09440

Wang H, Cheng S, Huang L, Guo W (2025) Investigating the meteorological causes of hydrological drought through the integration of spatiotemporal cubes and interpretable machine learning: A case study of the Yangtze River Basin. J Hydrol Reg Stud 62:102796. https://doi.org/10.1016/j.ejrh.2025.102796

Wanjiku P, Kennedy-Walker R (2024) Reviving Lake Victoria: A regional approach to inclusive sanitation. World Bank. https://www.worldbank.org/en/topic/water/publication/reviving-lake-victoria-a-regional-approach-to-inclusive-sanitation

Withers PJA, Jarvie HP (2008) Delivery and cycling of phosphorus in rivers: a review. Sci Total Environ 400(1–3):379–395. https://doi.org/10.1016/j.scitotenv.2008.08.002

Yuan Y, Zhou C, Wu J, Deng F, Liu W, Sun M, Li L (2025) An interpretable deep learning framework for river water quality prediction: a case study of the Poyang Lake Basin. Water 17(16):2496. https://doi.org/10.3390/w17162496

Zounemat-Kermani M, Alizamir M, Fadaee M, Sankaran Namboothiri A, Shiri J (2021) Online sequential extreme learning machine in river water quality (turbidity) prediction: a comparative study on different data mining approaches. Water Environ J 35(1):335–348. https://doi.org/10.1111/wej.12630

Acknowledgements

We gratefully acknowledge the invaluable contributions of the citizen scientists in Kenya and Tanzania, whose sustained commitment and careful observations made this work possible. These efforts were driven and coordinated through close collaboration with the Water Resources User Associations in Kenya and the Water User Associations in Tanzania, which played a central role in mobilizing participants, supporting field activities, and linking community-based monitoring with local water governance.

This project was supported by the World Water Quality Alliance Workstream on Citizen Science and the UNEP GEMS/Water Programme, whose guidance and support were instrumental in the development and implementation of this work.

We dedicate this work to the memory of Rochi Mkole, a key contributor to this project, whose enthusiasm, dedication, and hard work were central to many aspects of the study. Rochi passed away in January 2026, and his loss is deeply felt by the entire team. His commitment to community-based water quality monitoring continues to inspire this work.

Statements & Declarations

Funding

This research was funded by a grant from GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit). Additional support was provided through the UNEP SDG programme.

Data availability statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval*

This is not applicable.

Consent to Participate*

This is not applicable.

Consent to Publish*

This is not applicable.

Clinical trial number*

This is not applicable.

Study involves humans and/or animals*

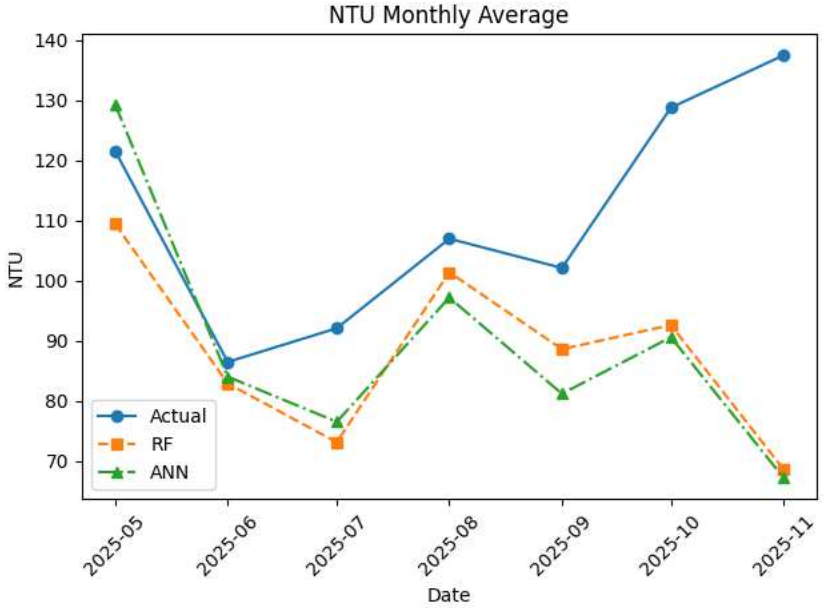
This is not applicable.

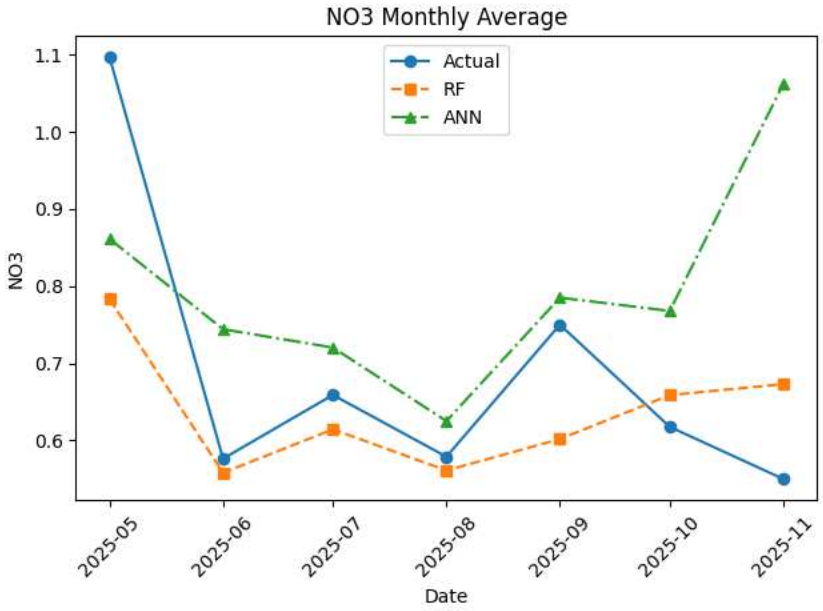
Authorship contribution statement

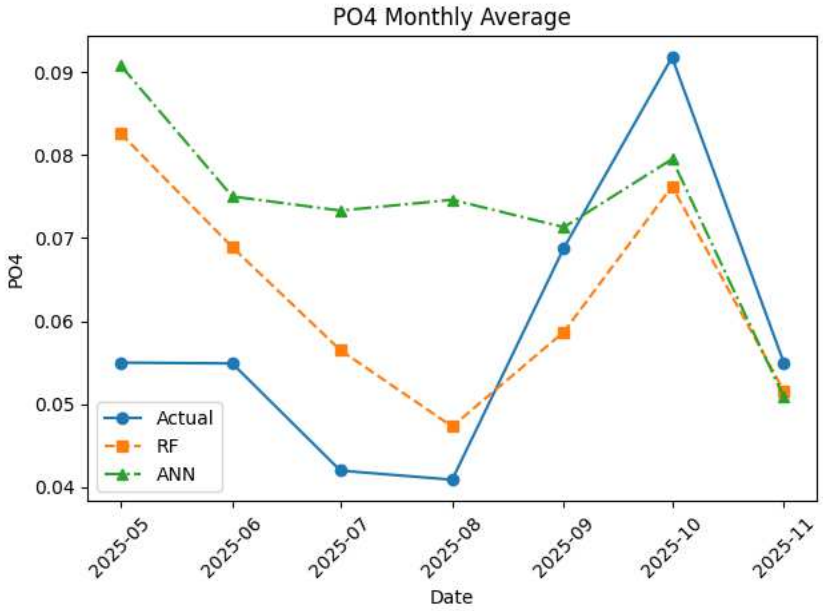
Xinyu Liu: Conceptualization**,** Writing-review, Data curation**,** and editing. **Georgio Emmanuel Nader:** Writing-review, Data curation**,** and editing. **Steven Loiselle:** Conceptualization**,** Writing-review, Formal analysis and editing. **Heri Chisute:** Writing-review and editing. **Fred Nyongesa:** Writing-review and editing. **Rochi Mkole:** Writing-review and editing. **Stuart Warner:** Writing-review and editing. **Luisa Galgani**: Writing-review and editing. **Amedeo Boldrini**: Writing-review. **Alessio Polvani**: Writing-review. **Riccardo Cirrone**: Writing-review.

Supplementary information

Supplemental Figure 1: **ANN and RF modelled water quality variables NTU, NO_3 and PO_4, averaged across all sub-basins, together with actual measured values for the independent test dataset.**







Supplementary Table 1. Comparison of plausible ANN architectures considered

Neural Network Type	Description	Strengths	Limitations for this Study	Suitability
Sequential ANN	Simple network where data flows from input → hidden → output layers	Works well with small & medium datasets, stable training, easy interpretation	No memory of temporal patterns	✓ **Chosen – most stable and appropriate**
Recurrent Neural Network (RNN)	Designed for sequential/temporal data	Captures time dependencies	Requires long sequences; dataset has irregular monthly sampling	✗ Not suitable
LSTM Networks	Advanced RNN variant	Good for long-term temporal patterns	Overly complex, prone to overfitting with small sample sizes	✗ Not suitable
Convolutional Neural Networks (CNN)	Extracts spatial features	Useful for raster imagery	Not suitable for tabular site-level data	✗ Not applicable

Hybrid CNN-LSTM Models	Spatial + temporal patterns	Effective for satellite time series	Requires a large dataset and spatial data grids	✗ Not appropriate
Multitask Neural Networks	Predict several targets jointly	Exploits shared structure between NO_3, PO_4, NTU	Requires larger balanced datasets	△ Potential but not optimal

Supplementary Table 2. Median concentrations of water quality indicators for each site (from upper basin to lower basin (6 to 1)).

Sub-basin Classification	Site	Count	NO_3 median	PO_4 median	NTU median
6	Bwawa La Manchira, Intake	13	0.1	0.01	14
6	Kisangura Darajani	8	0.1	0.01	70
6	Mto manchira makutano romakendo	10	1.125	0.1125	127
5	Upper Amala at Timoi	18	0.75	0.075	25
5	Isei river bridge D/S	20	1.5	0.055	100
5	Amala river upper bridge	29	0.75	0.035	100
5	Amala River at Mulot bridge D/S	39	0.75	0.035	150
4	Upper Nyangores at TMC bridge	10	0.1	0.01	23
4	Upper Nyangores at Lelbanga bridge	13	0.1	0.01	25
4	Nyangores River at Chemomul Bridge	33	0.35	0.035	35
4	Nyangores River at Bomet Town	33	0.75	0.035	40
4	Mto Rubana Darajani Makundusi	10	0.35	0.075	75
4	Illdungushu siana Wrua @ Nalengitomon	12	0.35	0.035	25
4	Talek river talek gate	15	0.35	0.035	150
3	Lower mara	38	0.75	0.075	100
3	Marasienta - Mara river	19	0.75	0.15	200
2	Mto Nyamonyo Mlimani Wegero	27	0.75	0.075	150
2	Mto Mara Nyamerambaro Kiriba	14	0.35	0.2125	14
2	Mto Nyarusondobiro Kwa Juma Magabe	31	0.35	0.035	200
2	Mto Mara Kitasakwa Kwiswa	23	0.35	0.075	40
2	Mto Nyamakendo Darajani	29	0.35	0.035	17
2	Mto tobora koreri	34	0.35	0.035	150

2	Mto Mara Kivuko	30	0.35	0.035	240
2	Mto Tobora Makutano	26	0.35	0.035	150
2	Mto Tigite Juu	37	0.35	0.035	75
2	Mto Tigite Chini Kewanja Darajani	31	0.75	0.01	75
2	Remaiso(North Mara Before Discharge)	30	0.75	0.035	75
2	Mto Nyaruherere Darajan Musati	30	0.35	0.035	240
2	Mto Somoche Darajani Kenyana	35	0.35	0.035	240
2	Nyantare(North Mara After Dischanging)	32	0.35	0.075	240
2	Mto Somoche Nyiboko	12	0.35	0.1125	240
2	Tigite Rechawi Matongo	34	75	0.75	0.035
2	Mto Nyamitita Darajani	22	0.35	0.075	57.5
2	Mto Somoche Nyamatare	25	0.35	0.01	240
2	Mto Mara, Bisarwi Burongo	16	0.75	0.075	20
1	Mto Mara Kirumi Darajani	15	0.35	0.035	100
1	Mto Mori -keisangora	16	3.5	0.15	27.5
1	Mto Mori Buganjo/Gamasara	14	0.35	0.035	100
1	Mto Mori, Utegi Darajani	18	0.35	0.035	200
1	Mto Mori Wamaya-Kirogo	15	0.35	0.075	75

Supplementary Table 3: Ranked feature importance for NTU, NO_3 and PO_4 prediction by RF, showing mean absolute SHAP importance and Spearman correlations (mean absolute SHAP values were used to assess overall feature importance)

Rank	Feature_NTU	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction
1	forest percent of sub-basin	24.810	0.244	0.332	+
2	precipitation	22.575	0.265	0.362	+
3	citizen scientist estimate of water flow velocity	11.985	0.221	0.262	+
4	Pasture extent of sub-basin	8.366	0.236	0.331	+
5	Soil Erosion of sub-basin	6.126	-0.029	-0.083	-
6	Road density of sub-basin	4.406	-0.131	-0.178	-
7	Human footprint of sub-basin 2009	3.354	0.020	0.005	≈0
8	Human footprints change of sub-basin 1993-2009	2.898	-0.014	-0.023	≈0
9	Population density of sub-basin	1.922	-0.116	-0.197	-
10	Urban extent of sub-basin	1.910	-0.114	-0.173	-
11	Night-time lights of sub-basin	1.598	0.011	-0.026	≈0

12	Population counts of sub-basin	1.475	-0.122	-0.202	-
13	cropland extent of sub-basin	1.428	0.085	0.074	+

Rank	Feature_NO3	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction
1	precipitation	0.137	0.037	0.007	≈0
2	Road density of sub-basin	0.081	0.296	0.348	+
3	Pasture extent of sub-basin	0.077	-0.306	-0.452	-
4	Human footprints change of sub-basin 1993-2009	0.066	0.093	0.040	≈0
5	citizen scientist estimate of water flow velocity	0.062	0.228	0.268	+
6	Population density of sub-basin	0.046	0.317	0.346	+
7	Urban extent of sub-basin	0.028	0.309	0.360	+
8	Soil Erosion of sub-basin	0.027	0.262	0.257	+
9	Population counts of sub-basin	0.026	0.338	0.399	+
10	Human footprint of sub-basin 2009	0.021	0.272	0.260	+
11	Night-time lights of sub-basin	0.019	0.115	0.102	+
12	cropland extent of sub-basin	0.015	0.193	0.225	+
13	forest percent of sub-basin	0.012	-0.257	-0.416	-

Rank	Feature_PO4	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction
1	precipitation	0.029	0.061	0.142	+
2	Pasture extent of sub-basin	0.013	0.047	0.219	+
3	citizen scientist estimate of water flow velocity	0.013	0.068	0.026	≈0
4	cropland extent of sub-basin	0.012	-0.098	-0.226	-
5	Population density of sub-basin	0.006	-0.092	-0.287	-
6	Soil Erosion of sub-basin	0.005	-0.004	-0.101	-
7	Population counts of sub-basin	0.004	-0.061	-0.238	-
8	Road density of sub-basin	0.004	0.051	-0.095	-
9	forest percent of sub-basin	0.004	0.020	0.163	+
10	Human footprint of sub-basin 2009	0.004	0.075	-0.043	≈0
11	Human footprints change of sub-basin 1993-2009	0.003	0.030	-0.097	-
12	Night-time lights of sub-basin	0.003	-0.061	-0.253	-
13	Urban extent of sub-basin	0.001	-0.058	-0.204	-

Supplementary Table 4. Ranked feature importance (absolute mean) for NTU, NO_3 and PO_4 prediction by ANN, showing mean absolute SHAP importance and Spearman correlations (mean absolute SHAP values were used to assess overall feature importance)

Rank	Feature_NTU	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction
1	forest percent of sub-basin	19.335	0.244	0.368	-
2	Urban extent of sub-basin	18.909	-0.114	0.164	-
3	citizen scientist estimate of water flow velocity	14.968	0.221	0.412	+
4	Human footprint of sub-basin 2009	14.583	0.085	0.260	+
5	Human footprints change of sub-basin 1993-2009	13.087	-0.014	0.456	+
6	cropland extent of sub-basin	13.053	0.265	0.314	+
7	precipitation	12.687	-0.189	-0.126	+
8	Population counts of sub-basin	10.911	-0.122	0.135	+
9	Soil Erosion of sub-basin	8.593	-0.131	0.167	-
10	Road density of sub-basin	7.979	-0.042	0.262	+
11	Pasture extent of sub-basin	6.891	0.020	0.389	+
12	Population density of sub-basin	3.272	-0.029	-0.094	-
13	Night-time lights of sub-basin	1.346	-0.116	0.174	-

Rank	Feature_ NO_3	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction
1	citizen scientist estimate of water flow velocity	0.224	0.272	0.319	-
2	Human footprint of sub-basin 2009	0.197	0.308	0.351	+
3	Pasture extent of sub-basin	0.194	-0.306	-0.497	+
4	Road density of sub-basin	0.163	0.338	0.514	+
5	Population counts of sub-basin	0.148	0.228	0.282	≈0
6	Population density of sub-basin	0.146	0.296	0.401	+
7	Soil Erosion of sub-basin	0.121	0.037	-0.104	+
8	precipitation	0.109	0.309	0.430	-
9	Urban extent of sub-basin	0.107	-0.257	-0.399	+
10	forest percent of sub-basin	0.088	0.262	0.235	≈0
11	Human footprints change of sub-basin 1993-2009	0.078	0.093	0.133	≈0
12	Night-time lights of sub-basin	0.063	0.317	0.479	-
13	cropland extent of sub-basin	0.052	0.193	0.327	+

Rank	Feature_ PO_4	SHAP importance mean	Spearman corr target	Spearman corr oof pred	Direction

1	Pasture extent of sub-basin	0.059	0.075	-0.061	+
2	Population density of sub-basin	0.032	0.102	0.012	-
3	Soil Erosion of sub-basin	0.030	-0.098	-0.240	≈0
4	forest percent of sub-basin	0.030	-0.061	-0.235	+
5	Urban extent of sub-basin	0.027	-0.061	-0.177	+
6	Human footprint of sub-basin 2009	0.021	-0.092	-0.282	-
7	Road density of sub-basin	0.020	0.030	-0.059	-
8	Population counts of sub-basin	0.019	0.047	0.171	+
9	cropland extent of sub-basin	0.016	-0.108	-0.218	≈0
10	citizen scientist estimate of water flow velocity	0.015	-0.058	-0.207	≈0
11	precipitation	0.014	0.068	-0.147	≈0
12	Night-time lights of sub-basin	0.012	0.020	0.128	≈0
13	Human footprints change of sub-basin 1993-2009	0.011	0.061	0.016	≈0

Supplementary Files

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

- ResponsetoReviewersorEditorsComments.docx