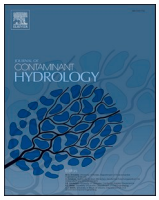
Journal of Contaminant Hydrology 251 (2022) 104078

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

journal homepage: www.elsevier.com/locate/jconhyd



Predicting in-stream water quality constituents at the watershed scale using machine learning

Itunu C. Adedeji [a], Ebrahim Ahmadisharaf [a,*], Yanshuo Sun [b]

[a] Department of Civil and Environmental Engineering, Resilient Infrastructure and Disaster Response Center, Florida A&M University—Florida State University College of Engineering, 2525 Pottsdamer St., Tallahassee, FL 32310, USA
[b] Department of Industrial and Manufacturing Engineering, Resilient Infrastructure and Disaster Response Center, Florida A&M University—Florida State University College of Engineering, 2525 Pottsdamer St., Tallahassee, FL 32310, USA

ARTICLE INFO

Keywords:
In-stream water quality
Machine learning
Seasonality
Uncertainty quantification

ABSTRACT

Predicting in-stream water quality is necessary to support the decision-making process of protecting healthy waterbodies and restoring impaired ones. Data-driven modeling is an efficient technique that can be used to support such efforts. Our objective was to determine if in-stream concentrations of contaminants, nutrients—total phosphorus (TP) and total nitrogen (TN) —total suspended solids (TSS), dissolved oxygen (DO), and fecal coliform bacteria (FC) can be predicted satisfactorily using machine learning (ML) algorithms based on publicly available datasets. To achieve this objective, we evaluated four modeling scenarios, differing in terms of the required inputs (i.e., publicly available datasets (e.g., land-use/land cover)), antecedent conditions, and additional in-stream water quality observations (e.g., pH and turbidity). We implemented five ML algorithms—Support Vector Machines, Random Forest (RF), eXtreme Gradient Boost (XGB), ensemble RF-XGB, and Artificial Neural Network (ANN) —and demonstrated our modeling framework in an inland stream—Bullfrog Creek, located near Tampa, Florida. The results showed that, while including additional water quality drivers improved overall model performance for all target constituents, TP, TN, DO, and TSS could still be predicted satisfactorily using only publicly available datasets (Nash-Sutcliffe efficiency [NSE] $>$ 0.75 and percent bias [PBIAS] $<$ 10%), whereas FC could not (NSE $<$ 0.49 and PBIAS $>$25%). Additionally, antecedent conditions slightly improved predictions and reduced the predictive uncertainty, particularly when paired with other water quality observations (6.9% increase in NSE for FC, and 2.7% for TP, TN, DO, and TSS). Also, comparable model performances of all water quality constituents in wet and dry seasons suggest minimal season-dependence of the predictions ($<$4% difference in NSE and $<$ 10% difference in PBIAS). Our developed modeling framework is generic and can serve as a complementary tool for monitoring and predicting in-stream water quality constituents.

1. Introduction

Elevated levels of in-stream pollutants are linked to water quality degradation and pose a significant hazard to human life and biodiversity (Alnahit et al., 2022). Despite efforts in water quality restoration over the years, ~41,000 waterbodies and 482,000 km of streams and shorelines are impaired nationwide as of 2012 (Copeland, 2012; Johnson et al., 2013). In 2022—50 years after the establishment of the Clean Water Act in 1972, this number has increased to over 1 million kilometers (~50% increase) for impaired rivers alone (Kelderman et al., 2022). Consequently, the average cost of developing and implementing Total Maximum Daily Loads (TMDLs) can be as high as ~$4.3 billion/year (USEPA, 2001).

Water quality management and restoration projects require adequate and continuous data for load reduction calculations (Borah et al., 2006, 2019; Mallya et al., 2020) and efficient modeling tools for timely water quality assessments. Non-point sources are the primary drivers of water quality degradation in many watersheds, and modeling in-stream pollution requires adequate assessments of these sources (Borah et al., 2006). Thus, the inter-linkage among environmental drivers such as watershed characteristics, meteorological, and water quality has been widely discussed in the literature (Cho et al., 2016; Fluke et al., 2019). While interactions among waterbody pollutants follow different linear and non-linear patterns constituting complexities in predictive

* Corresponding author.
E-mail addresses: cia16@fsu.edu (I.C. Adedeji), eahmadisharaf@eng.famu.fsu.edu (E. Ahmadisharaf), y.sun@eng.famu.fsu.edu (Y. Sun).

https://doi.org/10.1016/j.jconhyd.2022.104078
Received 17 May 2022; Received in revised form 9 September 2022; Accepted 11 September 2022
Available online 15 September 2022

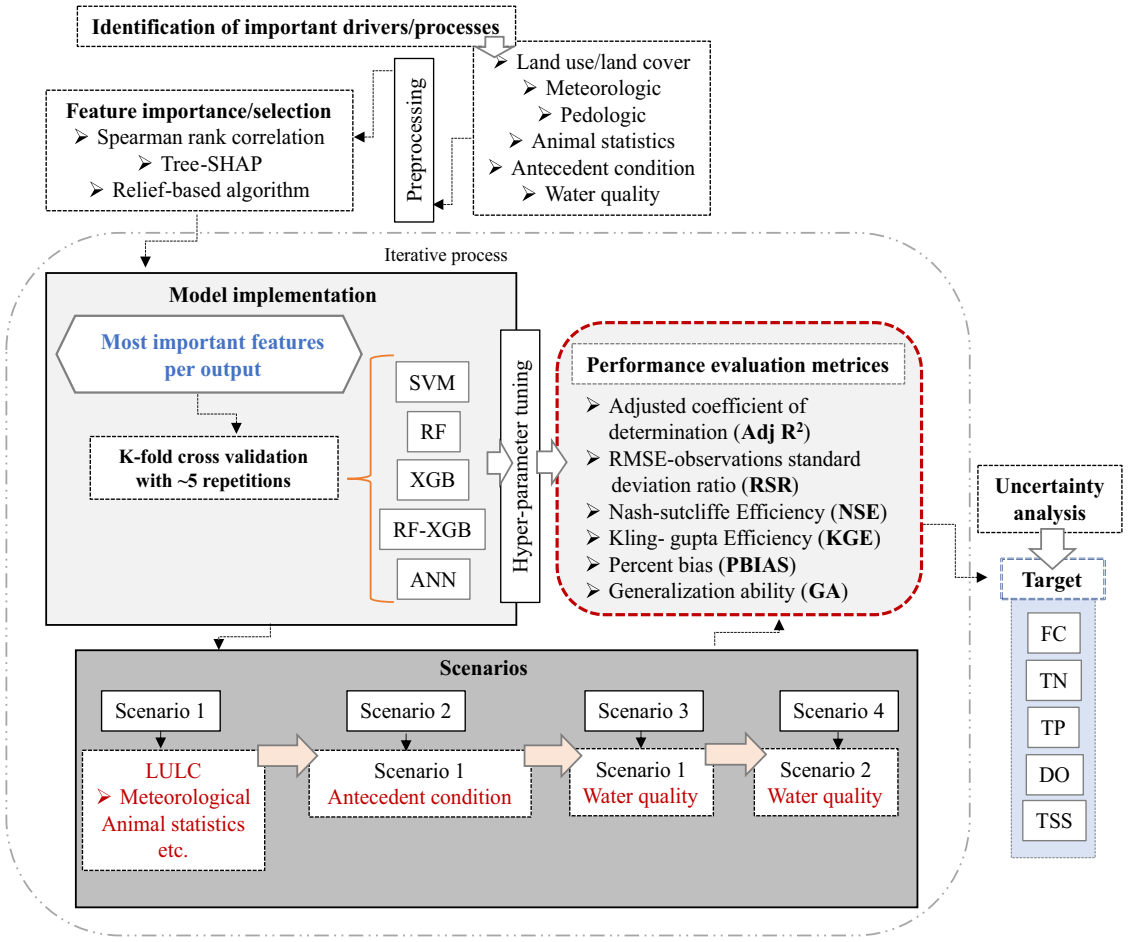


Fig. 1. Schematic of the machine learning-based water quality modeling framework detailing the modeling workflow starting from knowledge-guided feature selection to model performance evaluation and uncertainty analysis and uncertainty analysis. SVM: Support Vector Machines; RF: Random Forest; XGB: eXtreme Gradient Boosting; ANN: Artificial Neural Network; Tree-SHAP: Tree-based SHAPley Additive exPlanations; FC: Fecal coliform; TN: Total nitrogen; TP: Total phosphorus; DO: Dissolved oxygen; TSS: Total suspended solids.

modeling, sufficient or unavailable data further complicates these complexities.

Various combinations of environmental predictors have been employed in water quality modeling depending on data availability, water quality constituents of concern, and the scope of the study. Commonly used predictors include publicly available datasets such as streamflow, land cover, soil, meteorological (e.g., precipitation and air temperature), topography, and animal population, which explain underlying physical, chemical, and biological processes for water quality constituents (David and Haggard, 2011; Sakizadeh, 2016). Water quality observations are used to calibrate and validate water quality models (Khatri et al., 2020; Sakizadeh, 2016). Unlike hydrological data that are often obtained continuously, water quality observations are sparse due to the costs of monitoring and limited resources (Mallya et al., 2020). This is particularly the case for constituents like bacteria; the data are even sparser due to relatively more complex and expensive monitoring (Holcomb et al., 2018; Yu et al., 2021). Sparse datasets are some of the most significant challenges for modeling, especially for data-driven water quality modeling such as machine learning (ML) (Mallya et al., 2020a). Also, data quality comes into question; high concentration samples are crucial for pollution control studies like TMDLs, and their absence or inadequacy in datasets introduces bias to the water quality model predictions (Park and Engel, 2015). Furthermore, while some studies have successfully predicted pollutants without using other water quality constituents (Abimbola et al., 2020, 2021), other studies have suggested that their exclusion can lead to biased results; e.g., an overestimation reported by Park and Engel (2015) or underestimation found by Abimbola et al. (2021). Antecedent conditions with time windows (e.g., days) have been used to represent initial conditions, and studies have emphasized their importance in water quality modeling (Abimbola et al., 2020, 2021; Kao et al., 2020).

Adequate representation of in-stream water quality in watersheds requires an in-depth understanding of the underlying physical, chemical, and biological processes (Beven, 2018). Process-based models like Soil and Water Assessment Tool (SWAT) (Neitsch et al., 2011) and Hydrological Simulation Program Fortran (HSPF) (Bicknell et al., 2005) have been widely used in predicting water quality constituents at the watershed scale. These models are generally complicated and computationally demanding, particularly for large-scale watersheds and probabilistic analyses. They are also neither easy to implement, use, nor scalable. In situations of limited data, simple models, such as Load Estimator (LOADEST; Runkel et al., 2004), Web-based Load Interpolation Tool (LOADIN; Park and Engel, 2015), SPAtially Referenced Regressions On Watershed attributes (SPARROW; Schwarz et al., 2006), and load duration curves (Zhang and Quinn, 2019) have been used to generate water quality data and augment existing observations. These models are typically limited by many degrees of freedom and the assumption of linearity. In predicting water quality constituent loads using LOADEST, Park and Engel (2015) found significant bias in model predictions (Park and Engel, 2015). Their finding corroborates the study of Lee et al. (2016), who suggested that regression-based models can result in high systematic errors in conditions like heteroscedasticity of

model residuals, poor pollutant/flow correlation, and seasonality (Lee et al., 2016). In addition to the disadvantages mentioned above, these simple models do not account for pollution drivers (e.g., suspended solids and turbidity) and their underlying processes (e.g., settling and resuspension). These drivers and processes can be harnessed from data using statistical associations and dependencies in ML modeling (Wang et al., 2021).

Over the last decade, ML algorithms like ensemble ML (EML) and Artificial Neural Networks (ANN) have received attention in the area of water quality modeling (Abba et al., 2020; Al-Sulttani et al., 2021; Chen et al., 2020). This upward trend in the application of ML methods can be attributed to their ability to model intrinsic relationships in environmental systems while being computationally efficient, easy-to-use and easily automated (Abbas et al., 2021; Adams et al., 2013). ML algorithms can be used in conjunction with feature subset selection and model interpretation techniques, like SHapely Additive exPlanations (SHAP) and relief-based algorithms (RBA), to explain model outputs and quantify feature importance via a small number of water quality drivers (Bilali et al., 2021; Wang et al., 2021).

Feature selection varies widely in the literature due to constraints like data availability, spatiotemporal scale, and scope of the study. To the best of our knowledge, there are no systematic investigations into the importance of individual predictor sets, such as water quality drivers and antecedent conditions for watershed-scale water quality modeling. Additionally, model uncertainties associated with water quality predictions are rarely reported in ML-based water quality modeling studies (Duan et al., 2013; Farnham and Lall, 2015), posing a challenge for risk-based water quality management (Ahmadisharaf et al., 2019; Ahmadisharaf and Benham, 2020; Mishra et al., 2018, 2019). Thus, it is crucial to evaluate the effectiveness and uncertainty of ML-based modeling frameworks for in-stream water quality predictions using publicly available datasets alongside additional factors like antecedent conditions and other water quality constituents.

The main objective of this study was to evaluate the performance of five ML algorithms—Support Vector Machines (SVM), Random Forest (RF), eXtreme Gradient Boost (XGB), ensemble RF-XGB, and ANN—in predicting the in-stream concentration of fecal coliform (FC), nutrients—total phosphorus (TP) and total nitrogen (TN)—total suspended solids (TSS) and dissolved oxygen (DO) at the watershed scale using publicly available datasets. Through our analyses, we also aimed to: 1) determine if incorporating additional water quality constituents and accounting for antecedent conditions assist in improving the prediction of the five constituents; 2) investigate the role of seasonality on the predictive capability of ML models for the water quality constituents; and 3) quantify the uncertainty of ML predictions in terms of the water quality constituents. The model applicability was demonstrated on an inland stream, Bullfrog Creek Watershed, Tampa, Florida, where the abovementioned constituents are the primary sources of impairments, like many streams in Florida and the U.S.

2. Methodology

The modeling framework (Fig. 1) shows a sequence of steps, including identification of critical drivers (features or exploratory variables) of in-stream pollution in terms of the five water quality constituents, collection of the data representing pertinent processes, selection of the most important drivers as model features, implementation of ML algorithms in each scenario and evaluation of modeling scenarios using a set of fit metrics. Subsequently, seasonality and uncertainty analyses were conducted.

2.1. Exploratory data analyses and knowledge-guided feature selection

We conducted exploratory data analyses to investigate the relationships between the water quality constituents and their drivers. Here, "target" refers to a water quality constituent predicted by the ML algorithms, and "feature" is a variable that can explain the underlying physical, chemical, and biological processes of the target water quality constituent. Generally, increasing model dimension (number of features) exponentially increases the number of observations required (Tripathi and Govindaraju, 2007); hence, careful feature selection is essential to improve the model accuracy, and robustness and minimize the introduction of random errors from weak features into the learning algorithm (Kuhn and Johnson, 2019). Here, we selected the *"most important"* or *"most informative"* features that can explain the maximum target variability for each target water constituent.

We employed filter and embedded feature selection methods—spearman's rank-order correlation, tree-based Shapely Additive exPlanations (TreeSHAP), and relief-based algorithms (RBAs)—to extract critical features driving in-stream pollution from an initial set of 35 exploratory features. These methods use monotonic rank value (strongest monotonic association), RreliefF (Regression-based Relief) score (presence of feature value difference in adjacent data instances), and absolute shapely values (mean absolute SHAP value) in determining features with the highest importance. We used Spearman's correlation to prioritize the initial 35 features. This correlation is non-parametric and evaluates monotonic associations based on the rating of each variable, and as a result, linear and non-linear statistical dependencies among the target and features were assessed. One drawback is that spearman's ranking does not consider the dependencies between the target and features. Thus, we also used RreliefF—a relief-based algorithm—that does not assume independence among features, and is generally robust to missing data instances (Kononenko, 1994), high-dimensionality (Eiras-Franco et al., 2021), feature interactions (Urbanowicz et al., 2018), and noise in data (Eiras-Franco et al., 2021; Urbanowicz et al., 2018) in datasets. Lastly, we used TreeSHAP (XGB-SHAP)—a game theory-based method suitable for capturing complex feature-target relationships, robust to outliers, and provides interpretations for black box models like RF, XGB, and ANN. XGB-SHAP has been used in previous water quality prediction studies and found to be a highly effective model explanation method (e.g., Li et al., 2022). We repeated the aforementioned feature selection criteria (Spearman's correlation, RreliefF and XGB-SHAP) for each target water quality constituent and obtained a set of the *"most important features"* to be employed in the models. This multi-method approach was taken to obtain a credible set of explanatory variables for water quality modeling.

2.2. Machine learning algorithms

To identify the most suitable models for future water quality applications, five ML algorithms—SVM, RF, XGB, RF-XGB, and ANN—were selected and compared in this study. SVM and ANN were chosen due to numerous applications in previous environmental modeling studies (Banadkooki et al., 2020; Elshorbagy et al., 2005), while EML methods like RF and XGB are emerging algorithms in water quality applications (Li et al., 2022; Wang et al., 2021; Zounemat-Kermani et al., 2021). These methods are known for their adaptability to non-linear problems and remarkable modeling performance, as is the case in water quality modeling (Alqahtani et al., 2022; Li et al., 2022; Sakizadeh, 2016). In this study, we selected different EML methods (RF and XGB) based on the literature (Li et al., 2022; Wang et al., 2021; Zounemat-Kermani et al., 2021) and the training method employed. RF utilizes parallel training where each base learner is built independently using equal weights in each leaf node and eventual averaging all tree outputs. In the case of XGB, the base learners are generated sequentially, and a gradient is used from one learner to the others, thereby updating the leaf weights in subsequent trees. This way, each new tree corrects the error from the former giving more precise training predictions.

Although XGB is a stronger algorithm known for more precise training predictions over RF, it requires careful parameter tuning and is sensitive to overfitting in noisy data conditions (Zounemat-Kermani et al., 2021). The strength of XGB is an added advantage due reduction

Table 1
Model main functions and hyper-parameters.

Model	Estimators	Layers/NPL/ Iterations	Kernel	TF	Degree/ Gamma	C	Lr	Max depth[a]	Subsamples[b]	MC/CST/ MDS[c]	Bootstrap	MSS/ MSL[d]	ES[e]
SVM	–	–	rbf	–	3/0.25	1	–	–	–	–	–	–	–
RF	100	–	–	–	–	–	–	None	–	–	True	2/1	–
XGB	80	–	–	–	–	–	0.1	4	1	1/1/0	–	–	10
RF-XGB	100/80	–	–	–	–	–	—/0.1	None/4	—/1	—/1/1/0	True/—	2/1/—	—/10
ANN	–	2/20/20	–	Tansig	–	–	0.001	–	–	–	–	–	–

SVM: Support Vector Machine, RF: Random Forest; XGB: eXtreme Gradient Boosting, ANN: Artificial neural network. NPL: Neurons per layer, TF: Transfer function, Degree/Gamma: The coefficient of Radial basis function kernel that specifies the minimum loss reduction required to make a split, C: Regularization parameter, Lr (Learning rate/eta): The step size shrinkage used to prevent overfitting of the model.
[a] Maximum depth of a tree: Used to control overfitting.
[b] Fraction of observation to be randomly sampled for each tree.
[c] MC/CST/MDS: Minimum child weight/Columns sample by tree /Maximum delta step.
[d] MSS/MSL: Minimum sample split/Minimum sample leaF.
[e] ES: Early stopping round; rbf: Radial basis function kernel.

in error propagation when building subsequent base learners, while overfitting occurs when the model memorizes random errors (noise) in the training data. On the other hand, RF is intrinsically robust to overfitting due to the random selection of inputs but generally requires higher computational time than XGB (Stocker et al., 2022; Zounemat-Kermani et al., 2021). In an effort to minimize the limitations and leverage the strengths of individual algorithms, we implemented RF-XGB using Voting Regressor (a sci-kit learn package). This approach was supported by Zounemat-Kermani et al. (2021), who suggested that combining multiple ML algorithms through model averaging can leverage the strengths of individual algorithms and improve generalization ability (Zounemat-Kermani et al., 2021).

2.3. Performance evaluation metrics for machine learning models

Parameters of each selected ML algorithm were optimized, and their predictive performances were evaluated. Employing multiple fit metrics is helpful in modeling studies as the strengths and limitations of individual measures can be leveraged (Ahmadisharaf et al., 2019). Here, we used six metrics to evaluate the performance of the ML algorithms in terms of bias, error, and correlation. The metrics included the adjusted coefficient of determination (Adj R^2), Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), percent bias (PBIAS), root mean square error (RMSE)-observations standard deviation ratio (RSR), and Kling-Gupta efficiency (KGE; Gupta et al., 2009). Based on these metrics, the performance of each ML algorithm was ranked using suggested guidelines for accurate quantification in watershed modeling (Ahmadisharaf et al., 2022; Ahmadisharaf et al., 2019; Lamontagne et al., 2020; Moriasi et al., 2015). In addition to these metrics, we used generalization ability (GA) to ensure each model performed satisfactorily with the introduction of new data (Bilali et al., 2021). A GA value of unity (1) indicates excellent model generalization, GA $>$ 1 indicates overfitting, while values less than one ($<$1) indicates an underfitting (Bilali et al., 2021). The term '*improvement in model performance*' in this study refers to a significant improvement recorded in one or more of the selected objective functions. Likewise, the term '*best model*' used in each modeling scenario was selected based on larger Adj R^2, NSE, and KGE, and lower RSR and PBIAS. Hereafter, "*excellent*", "*very good*", "*good*", and "*satisfactory*" are derived based on Moriasi et al. (2015) and Ahmadisharaf et al. (2019). We computed these metrics using Hydrostats and HydroErr packages in Python. The reader is referred to the papers mentioned above for details of the selected evaluation matrices.

2.4. Training and testing of machine learning models

The ML algorithms were evaluated using a cross-validation (CV) technique with multiple repetitions (typically five) to assure valid prediction results due to the learning algorithm's stochastic nature, i.e., samples are randomly split in each learning scenario. Although 5-fold CV (i.e., using 80% of observation for training and the remaining 20% for testing) is typically found suitable in many modeling practices, we examined 3-, 4- and 7-fold CV train/test ratios alongside the 5-fold for all the ML algorithms. Our rationale was that the ML model performance increases with data volume, so the ratio of the training and hold-out sets was kept constant across the learning scenarios for each target water quality constituent to ensure the impartial model and scenario comparison. Unfitted RF and XGB models were combined using an ensemble meta-estimator (voting regressor in Scikit-learn Python library) that averages the predictions of each estimator. Also, since adjusting the model hyper-parameter (factors that regulate the learning process) manually is inefficient due to numerous tunable functions (Table 1), we used Levenberg-Marquardt and early stopping optimization techniques to perform an exhaustive search and determine the optimum value of the model hyperparameters. Levenberg-Marquardt technique, a simple and robust function approximation method, was used to optimize ANN model performance, while SVM, RF, and XGB were optimized using grid search tuning technique and early stopping. Also, due to the size of our data—234 for FC, TN, TP, and DO, and 80 for TSS—the learning algorithms chosen in this study tended to overfit the observed data; therefore, to minimize the variance, Levenberg-Marquardt and early stopping optimization were used in the ANN and tree-based algorithms respectively. While a step size of 0.001 yielded an optimal performance in the neural network, increasing either iterations or neuron count above 20 did not improve the learning performance. Table 1 lists the primary hyper-parameters and functions for all the ML algorithms.

Table 2
Machine learning-based water quality modeling scenarios.

Scenario	Drivers	Description
S1	LULC, meteorological, pedologic, animal statistics	Evaluates the suitability of using only publicly available dataset.
S2	LULC, meteorological, pedologic, animal statistics, antecedent condition	Evaluate the importance of accounting for antecedent conditions.
S3	LULC, meteorological, pedologic, animal statistics, water quality	Evaluate the importance of accounting for other in-stream water quality constituents.
S4	LULC, meteorological, pedologic, animal statistics, antecedent condition, water quality	Evaluate the importance of accounting for antecedent conditions and other in-stream water quality constituents (combination of scenarios S1-S3).

LULC = Land use/Land cover.

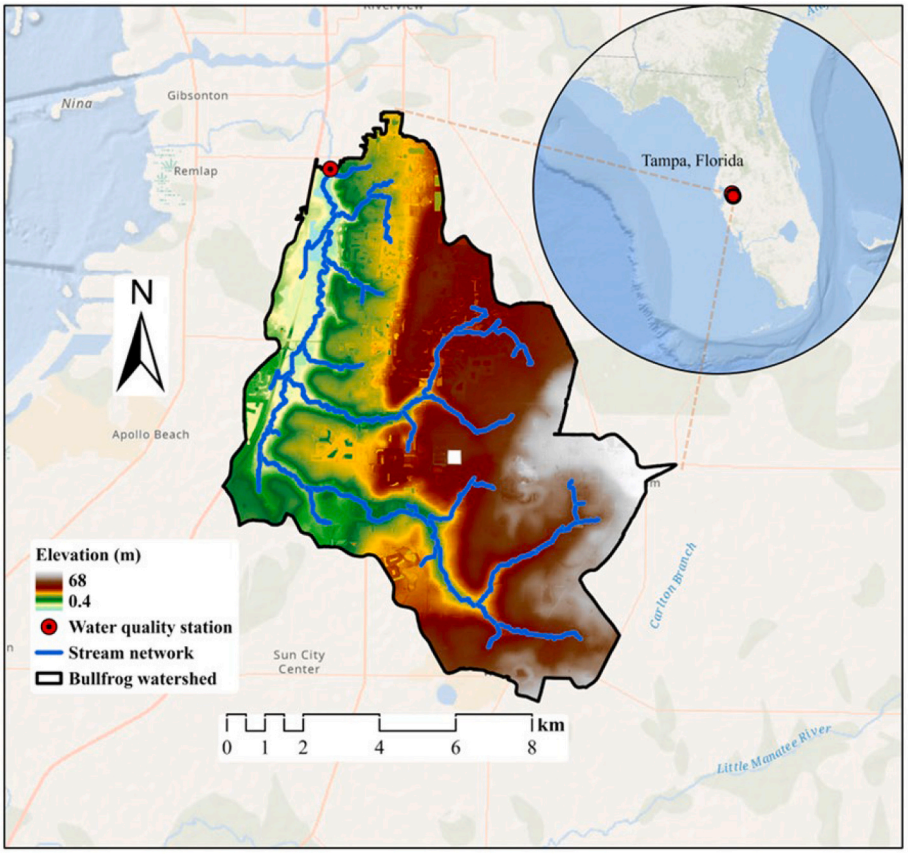


Fig. 2. Bullfrog River Watershed.

2.5. Modeling scenarios

Modeling scenarios (Table 2) were designed such that each scenario represents the water quality processes and data requirement considerations. In Scenario 1 (S1), we assessed only the following variables: land use/land cover (LULC), meteorological, soil animal statistics, etc. whose data are publicly available throughout the U.S. In scenario 2 (S2), we considered the role of antecedent conditions, which can influence contaminant leaching, water seepage, and overland flow volume. Different representations of antecedent conditions have been incorporated in previous studies, such as cumulative rain prior to a water quality observation, antecedent dry days, weighted antecedent rainfall, or a combination of different factors in one variable (Abimbola et al., 2020; Dada and Hamilton, 2016; Farnham and Lall, 2015). Thus, we evaluated cumulative rainfall over a range of periods (one to 14 days), antecedent dry days, and different sets of three-day weight-adjusted prior to an observed water quality constituent as representations of the antecedent conditions.

These conditions can influence in-stream pollutant dynamics; accumulated pollutants on the land surface can be mobilized and conveyed during a rainfall event and subsequently transported to receiving waters via succeeding rainfall events (wash-off). Many antecedent rainfalls can pronounce the effect of rainfall on the day of observation due to increased soil moisture content, decreased soil infiltration capacity due to saturation, and increased surface runoff (Farnham and Lall, 2015). Furthermore, in scenario 3 (S3), we evaluated the effect of incorporating additional water quality constituents (e.g., pH and turbidity) to represent in-stream water quality processes like settling/resuspension and die-off/regrowth without considering antecedent conditions. Since data on these additional water quality constituents may be proprietary, limited, or unavailable in many watersheds, this scenario is applicable only when such data are available, or resources exist to collect them through monitoring. Finally, Scenario 4 (S4) combines scenarios S1-S3, representing a situation where all data mentioned above are available.

2.6. Uncertainty quantification

The uncertainty of ML algorithms was quantified using a robust estimator of variance suggested by Wahl (2004) and Ahmadisharaf et al. (2016). In this approach, the uncertainty of each ML algorithm is presented in terms of a prediction interval around a hypothetical value of unity. The overall methodological sequence includes an initial computation of prediction errors (Eq. (1)), exclusion of outliers from the prediction errors series, estimation of the mean and standard deviation of the prediction errors, and determination of the confidence band around the predicted values of each target water quality constituent (Eq. (2)).

$$Prediction\ error \quad (e) = Log_{10}\left(Y^{sim}\right) - Log_{10}\left(Y^{obs}\right) \tag{1}$$

$$Prediction\ interval\ band \quad [Upper\ limit, Lower\ limit] = \left[Y^{sim} * 10^{-e^{mean}-2S_e}, Y^{sim} * 10^{-e^{mean}+2S_e}\right] \tag{2}$$

where, e is the prediction error, e^{mean} and S_e are the mean and standard deviation of the prediction error, respectively and $\pm\ 2S_e$ provides the 95% prediction interval. The estimated prediction interval for each water quality constituent can be used as multipliers to achieve a range of predictions when using the ML models in other water quality modeling scenarios.

2.7. Seasonality analyses

Seasonal influence on prediction performance is essential, mainly when profound seasonal trends are present in the observations. High intensity and more frequent precipitations are prevalent in the wet

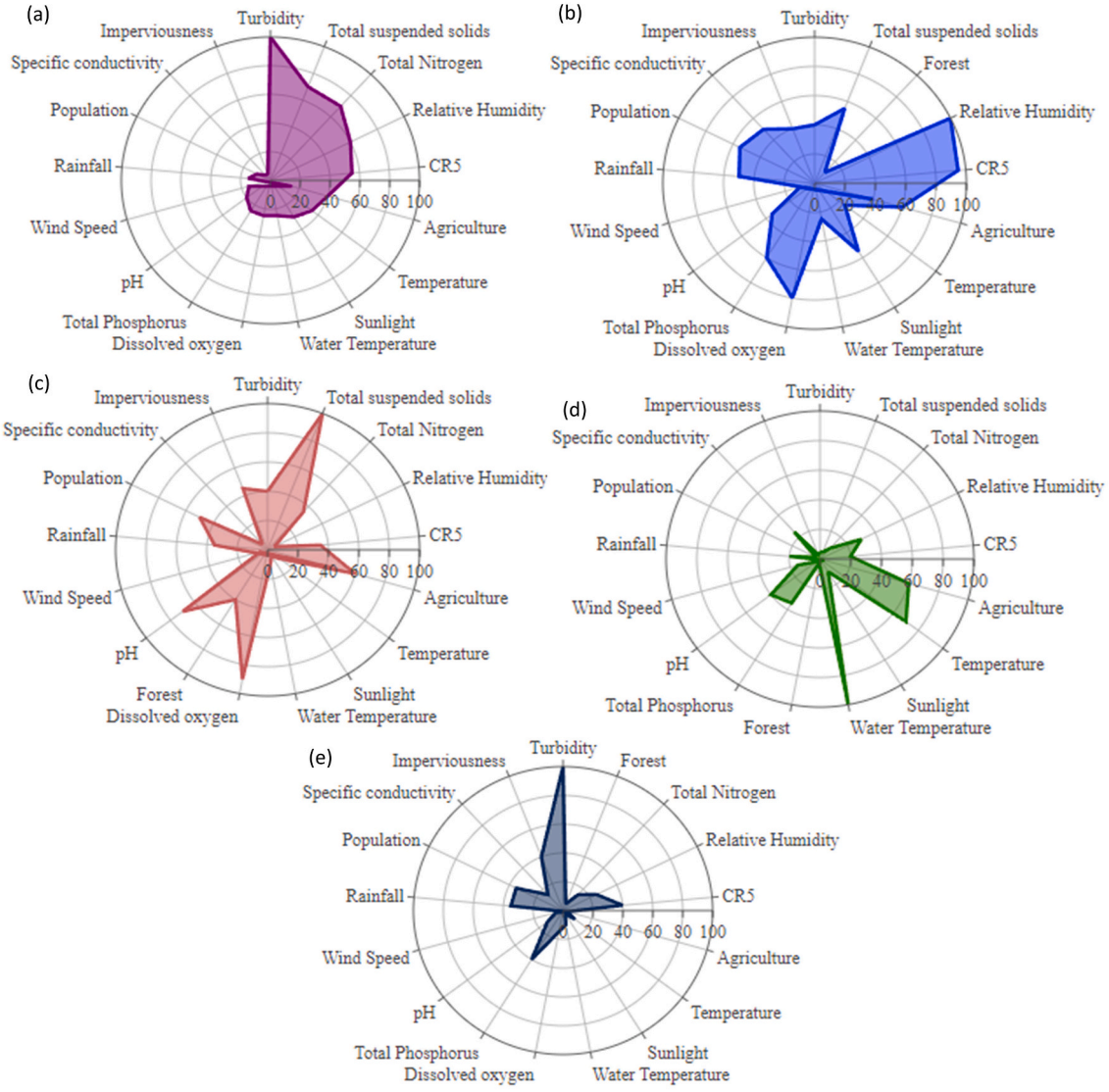


Fig. 3. Global feature importance for: (a) fecal coliform (FC); (b) total nitrogen (TN); (c) total phosphorus (TP); (d) dissolved oxygen (DO); and (e) total suspended solids (TSS), assessed by the combination of three feature importance methods (XGB-SHAP, ReliefF [a relief-based algorithm] and Spearman correlation). The influence of water quality constituents is more pronounced for FC, TP, TSS, and TN, while metrological variables dominate for DO. CR5: 5-day antecedent cumulative rainfall, XGB-SHAP: eXtreme Gradient Boosting Trees-SHapely Additive exPlanations.

season, which increases the overland flow and transport of contaminants to the stream through wash-off processes (Moriasi et al., 2014; Sigleo and Frick, 2003). In a dry season, sporadic short and high-intensity rainfall events are prevalent, promoting the transport of contaminants to adjacent streams and tributaries. Here, we developed the ML algorithms for wet and dry seasons following the methodology presented in Sections 2.1–2.6. We then compared the performance of these algorithms to investigate whether the performance is season dependent.

3. Case study

The study area is the freshwater segment of the Bullfrog River Watershed located in the southern region of Hillsborough County, Tampa, Florida (Fig. 2). The watershed covers a drainage area of 104.2 km^2, corresponding to the Hydrologic Code Unit (HUC031002060401) according to the U.S Geological Survey classifications. The climate in this watershed is humid subtropical, where air temperatures range from daily lows of 8 °C to as high as 34 °C. The watershed has a heterogeneous landscape with developed, agriculture, wetlands, open waters, barren land, and forest land covers. Although historically, the watershed was considered an agricultural watershed dominated by pasturelands, over the years, there has been substantial urban sprawl across the downstream, which is the area closest to Tampa metropolitan area. Urban is the dominant land cover in the watershed downstream, while the upstream part is still predominantly agricultural (pasture and cropland). The soil is mainly composed of highly permeable soil, with a blend of hydrologic soil groups A and A/D. The watershed is low-lying (average ground slope of 2.5%) with an average water table depth of 0 to 1.45 m. Additional information on data sources and temporal coverage can be found in the supplementary information document (Table S1).

4. Results and discussion

4.1. Statistical analyses and feature importance

Log transformations were applied to all target time series to remove skewness in the distributions and deemphasize outliers, thereby improving statistical validity of the observations for our predictions.

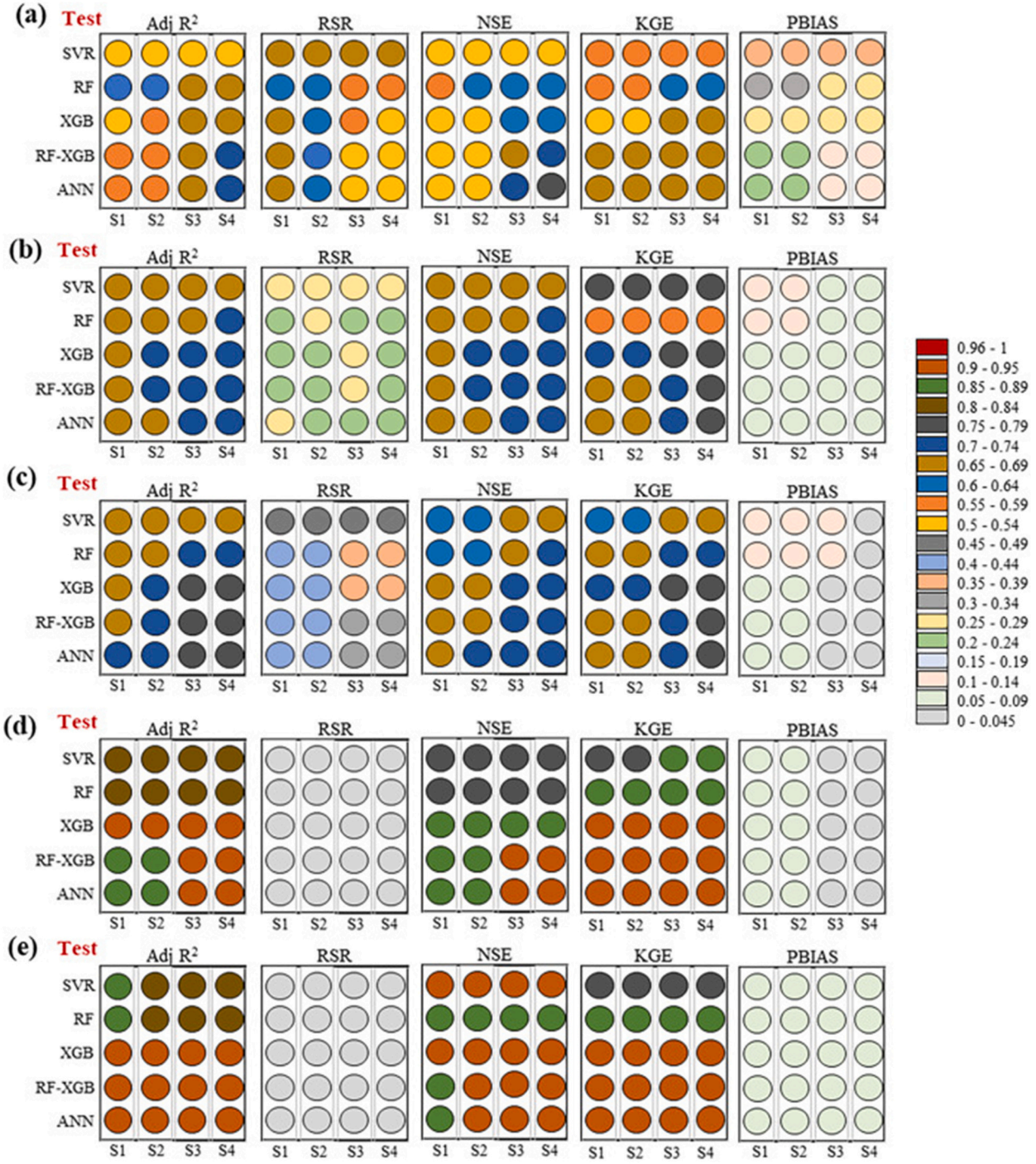


Fig. 4. Machine learning model performances—coefficient of determination (R^2), root mean square error-observations standard deviation ratio (RSR), Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE), and percent bias (PBIAS)—in the test phases of the four modeling scenarios (S1-S4) for: (a) fecal coliform (FC), (b) total nitrogen (TN), (c) total phosphorus (TP); (d) dissolved oxygen (DO); and (e) total suspended solids (TSS). The predictions of TN, TP, TSS, and DO were satisfactory in all modeling scenarios. SVR: Support Vector Machines; RF: Random Forest; XGB: eXtreme Gradient Boosting Trees; RF-XGB: Coupled RF and XGB; ANN: Artificial Neural Network.

Descriptive statistics such as minimum, maximum, mean, and quantiles of log-transformed target constituents are shown in Fig. S3. Additional analyses of the time series show that the distribution of rainfall, five-day antecedent rainfall (CR5), FC, DO, and TSS had no significant trend from 1998 to 2016. Conversely, TP had a significant decreasing trend (p-value = 0) whereas a slightly increasing trend was observed for TN (p-value = 0.02).

Furthermore, seventeen '*most important*' features were selected and implemented in the ML algorithms. We examined the cumulative rainfall over a range of periods (one to 14 days), antecedent dry days, and different sets of 3-day weight-adjusted prior to an observed water quality constituent to represent the most crucial antecedent conditions in terms of each of the five water quality constituents. Our results showed that CR5 was more important than other antecedent condition features for all the water quality constituents. Subsequently, we obtained a union feature set of the top 17 features after combining the

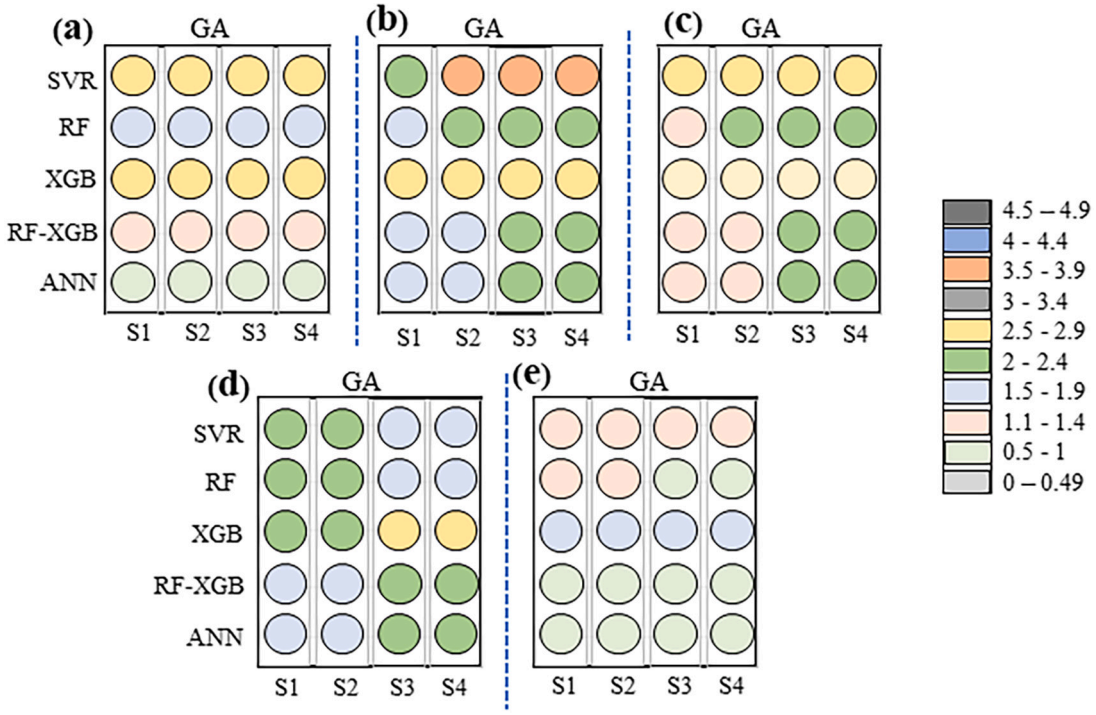


Fig. 5. Generalization ability (GA) of the machine learning algorithms in the four modeling scenarios (S1 - S4) in terms of generalization ability (GA) during training and test phases (combined) for: (a) fecal coliform (FC); (b) total nitrogen (TN); (c) total phosphorus (TP); (d) dissolved oxygen (DO); and (e) total suspended solids (TSS). Rf-XGB and ANN were the best generalizable models, while DO and TSS had the best GA values. SVR: Support Vector Machines; RF: Random Forest; XGB: eXtreme Gradient Boosting Trees; RF-XGB: Coupled RF and XGB. ANN: Artificial Neural Network.

results from features importance analyses of each target water quality constituent. Fig. 3 shows that the final selection of '*most important* features' (top 17 features) was consistent across target variables, except for the inclusion of 'forest area' in predicting TN, TP, DO and TSS. Also, meteorological variables were more significant for TN and DO, while the inclusion of additional water quality constituents (Scenarios S3 and S4) dominated in predicting FC, TP, and TSS. Agricultural area (combined pasture, cropland, and grassland), forest area, and human population were the most important factors in predicting FC, TN, and DO, while surface imperviousness was more important for TP and TSS predictions. Similarly, antecedent conditions were more important than rainfall for FC, TP, and TSS, further highlighting the importance of accounting for infiltration/runoff effect and soil memory in watershed-scale water quality predictions.

Despite the low-lying characteristic of the study watershed, baseflow was not found to be an important feature and was excluded from the final feature set. Likewise, animal population was excluded due to its relatively weaker importance than other features, which can stem from significant temporal and spatial approximation in the data. Other excluded variables included soil moisture, wind speed, atmospheric pressure, and vapor pressure deficit. It is important to note that limiting factors can present unavoidable bias in our modeling framework. Such factors include but are not limited to insufficient temporal coverage for some features and inadequate information on others, including septic tanks, wildlife population, animal grazing patterns, proximity to the stream, and fertilizer application.

4.2. Model performance

In this section, we evaluated the selected methods in terms of their overall effectiveness in predicting target water quality constituents for each modeling scenario. Fig. 4 and Fig. 5 detail the performance of each model in terms of the five water quality constituents using the six model fit metrics while Fig. 6 reveals the average percent improvement from scenarios S1 to S4. The reader is referred to Fig. S1 for information on training performance and Fig. S2 for Quantile-Quantile probability plots of observed and predicted values in all scenarios.

4.2.1. Fecal coliform (FC)

We found that FC cannot be satisfactorily by solely using publicly available datasets (NSE = 0.5 and PBIAS = 24.5 in Scenario S1). However, the performance of ML algorithms improved from Scenario S1 to S4, suggesting that satisfactorily prediction of this water quality constituent can only be achieved when additional water quality constituents (turbidity, TN, TP, water temperature, DO, pH, specific conductivity, and TSS) are used along the publicly available datasets. Overall, there was a 1.8%, 24.0%, and 30.9% improvement in R^2 in Scenarios S2, S3, and S4 (Fig. 6). PBIAS values averaged between 24.5% (S1) and 15.5% (S3) for most ML models, while ANN had the lowest PBIAS (13.4%) in Scenario S4. According to RSR values in Fig. 4a, prediction errors decreased considerably from Scenario S1 to S4 (20.1% and 22.2%) reduction in RSR for training and test phases, respectively, with RF-XGB and ANN having the lowest values in Scenario S4 (0.48 and 0.49, respectively). In Scenarios S1 and S2, error values were generally satisfactory ($0.6 < RSR < 0.7$), and good in Scenarios S3 and S4 ($0 < RSR < 0.5$). Similarly, GA values indicated good generalizability without significant overtraining across all the modeling scenarios (~1.4 on average), except for SVR with an average GA of 2.6 (Fig. 5).

4.2.2. Total nitrogen (TN) and total phosphorous (TP)

TN and TP predictions were satisfactory and similar across the four modeling scenarios, suggesting that these nutrients can be predicted satisfactorily by solely using the publicly available data (NSE = 0.68 and 0.69 for TN and TP, respectively). The prediction accuracy, however, improved from Scenario S1 to S2, with more improvement evident for TP (NSE = 0.69 and 0.73 for TN and TP, respectively) in the test phases, as shown in (Fig. 6, Fig. 7b and Fig. 7c). Likewise, prediction accuracy was significantly improved in Scenario S3 (17.0% for TP and 14.8% for

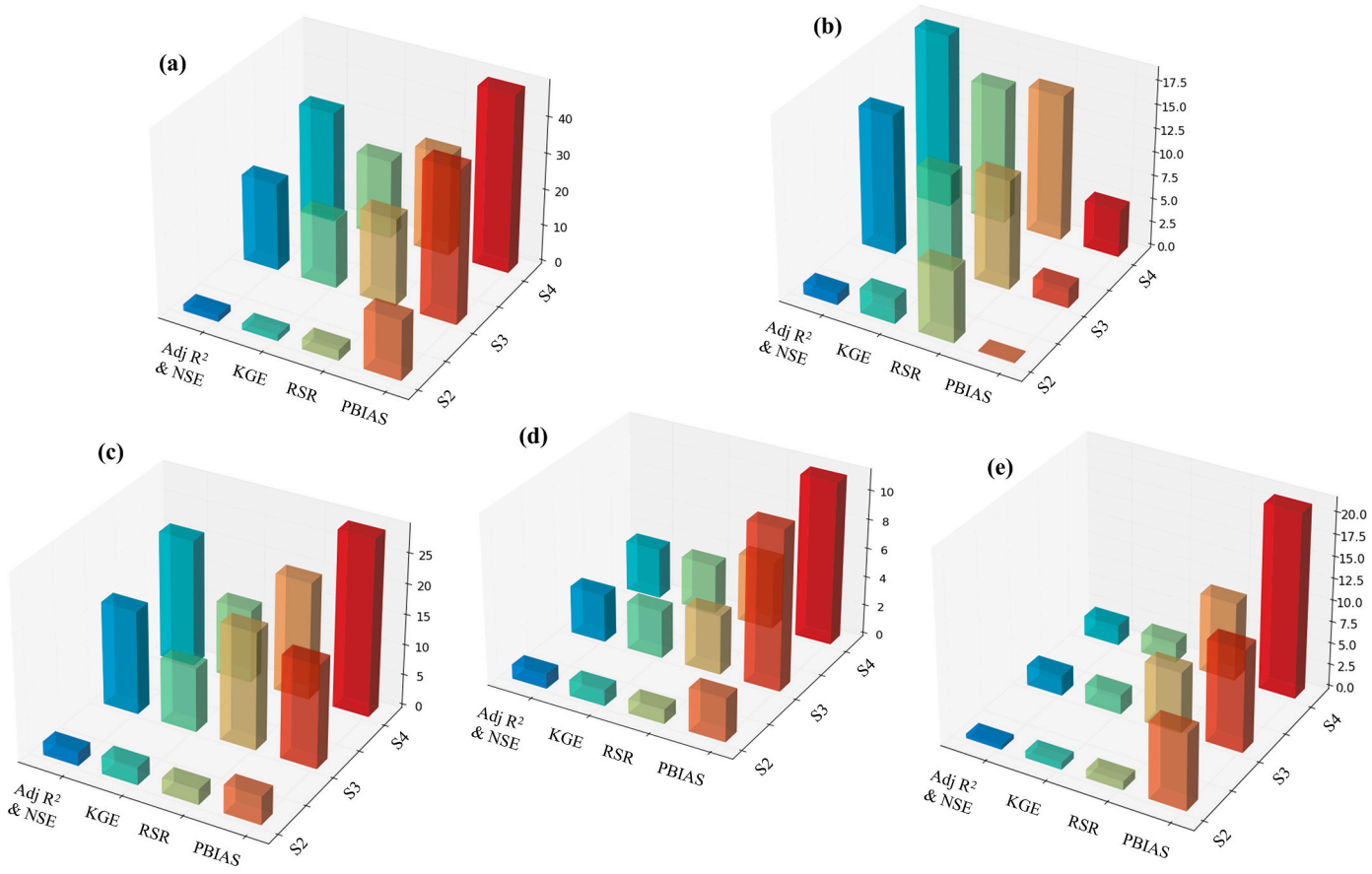


Fig. 6. Average percent improvement (%) of the machine learning algorithms in terms of Adjusted R2 (Adj R2), Nash Sutcliffe Efficiency (NSE), Kling Gupta Efficiency (KGE), root mean square error (RMSE)-observations standard deviation ratio (RSR), and Percent bias (PBIAS) for modeling scenarios S2, S3 and S4 compared to scenario S1 (solely using publicly available data).

TN) and peaked in Scenario S4 (21.1% for TP and 18.5% for TN). As observed for FC, the performance improvement from Scenario S3 to S4 was more prominent than that from Scenario S1 to S2, suggesting that the influence of antecedent conditions was more significant when combined with additional water quality constituents (turbidity, water temperature, dissolved oxygen, pH, specific conductivity, and total suspended solids).

4.2.3. Dissolved oxygen (DO) and total suspended solids (TSS)

The predictions of DO and TSS were very good ($R^2 > 0.7$) in all the modeling scenarios. Though there were no significant differences in the training accuracies of DO and TSS predictions for all the ML models, PBIAS was greatly improved in Scenario S4, particularly for TSS (11% in DO and 21.2% in TSS; Fig. 6). XGB and RF-XGB slightly outperformed other models in the test phases of DO predictions.

4.2.4. Fecal coliform (FC), total nitrogen (TN), total phosphorous (TP), dissolved oxygen (DO), and total suspended solids (TSS)

While the RSR values of TN, TP, DO, and TSS were minimal in all the modeling scenarios, there was a notable decrease in RSR values in scenarios S3 and S4 from S1 (average values of 17.4 in S3 and 18.6 in S4; Fig. 6). This decrease was more pronounced for FC, TP, DO, and TSS than other target variables. Even though TP was better predicted than TN, TN predictions had smaller RSR values than TP. This can be due to a larger standard deviation of the TN observations. Furthermore, the average PBIAS across all modeling scenarios were marginal in the testing phase, with values of 8.7%, 5.5%, 2.2%, and 8.4% for TN, TP, DO, and TSS, respectively. A closer look at PBIAS variability across all the target constituents revealed a similar pattern as seen in the prediction uncertainty, where FC, TSS, and TN had the largest PBIAS. For TN, TP, and DO, model overfitting reduced was as indicated by the decrease in GA value in Scenarios S3 and S4 (Fig. 7). Decreasing GA value in the same experiment can be localized to a larger training error suggesting that the prediction improvement (decrease in variance) observed in these scenarios was more prominent in the test phase than the training phase. Conversely, unity values of GA in all the modeling scenarios were observed for TSS, suggesting that TSS predictions have more generalizability than all other constituents. Furthermore, although SVM had significant overfitting for FC, TN, and TP (average GA = 3.4), average values of GA were excellent in predicting TSS (GA = 1), while DO's prediction had a slight overfitting problem DO (GA = 1.7). Despite considerable effort to minimize variance in test predictions, considerable overfitting was observed in the predicted results of the XGB model across all scenarios (GA = 2.7 for FC, GA = 3.7 for TN, GA = 2.6 for TP, GA = 2 for DO, and GA = 1.3 for TSS; Fig. 5). On the other hand, RF-XGB presented the most generalizable predictions (GA = 1.2 for FC, GA = 2 for TN, GA = 1.5 for TP, GA = 1.9 for DO, and GA = 1 for TSS; Fig. 5) across all modeling scenarios compared to RF (GA = 1.7 for FC, GA = 2.4 for TN, GA = 2.2 for TP, GA = 2 for DO, and GA = 1.5 for TSS; Fig. 5). Based on average GA values for FC, TP, TN, DO, and TSS (across scenarios), RF-XGB provided a 36% and 22.4% reduction in model overfitting for XGB and RF-XGB models, respectively. In conclusion, among the five ML models, ANN and RF-XGB had the lowest overfitting problems (Fig. 5).

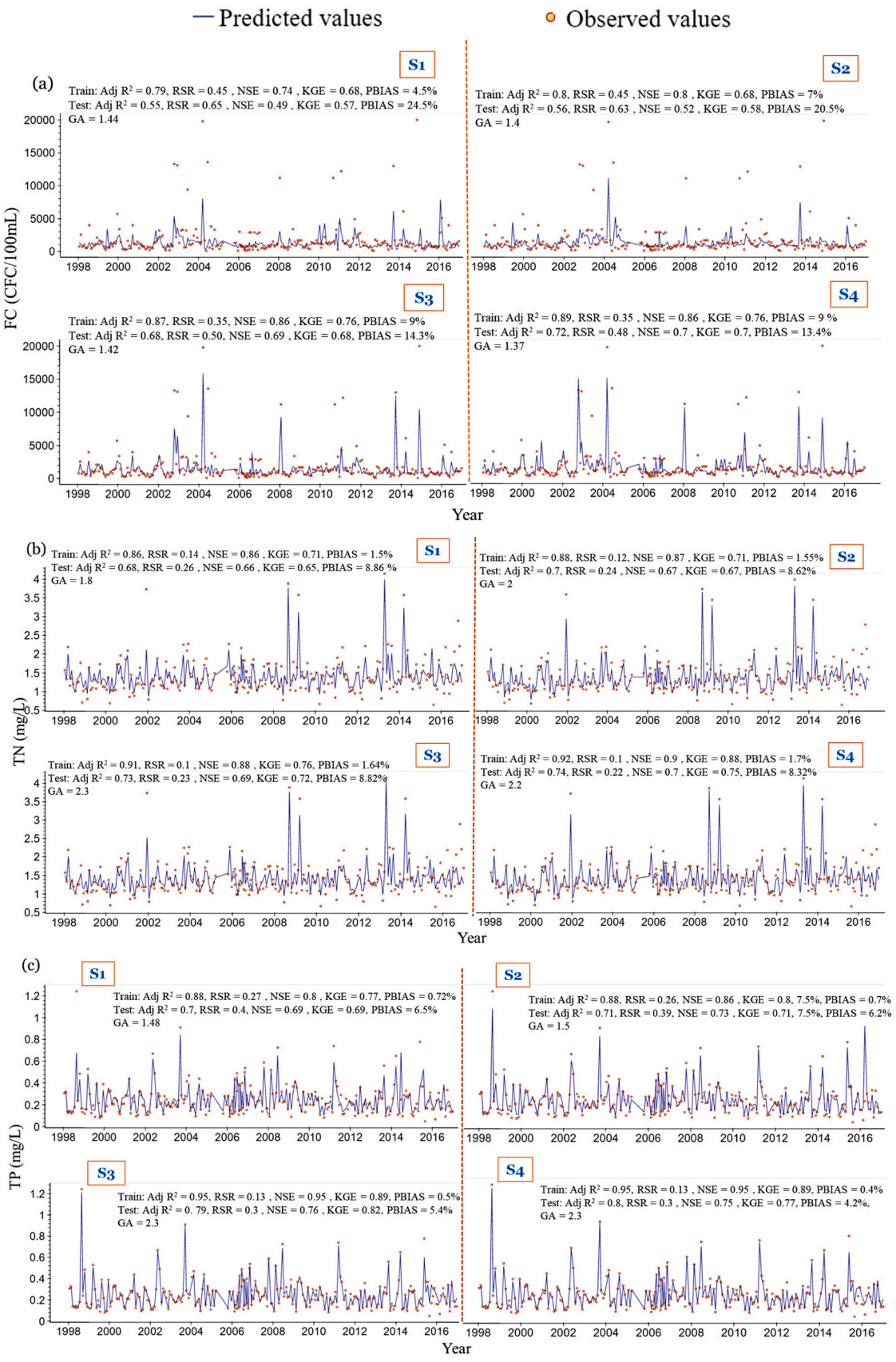


Fig. 7. Observed and predicted instantaneous time series of: (a) fecal coliform (FC); (b) total nitrogen (TN); (c) total phosphorus (TP); (d) dissolved oxygen (DO); and (e) total suspended solids (TSS) showing performance of the *"best-performing"* machine learning algorithm in the four modeling scenarios (S1 - S4). Overall, the modeling scenario S4 led to a better match with the observations. This is evident for FC, where high concentrations are better captured in this modeling scenario (introducing additional water quality constituents).

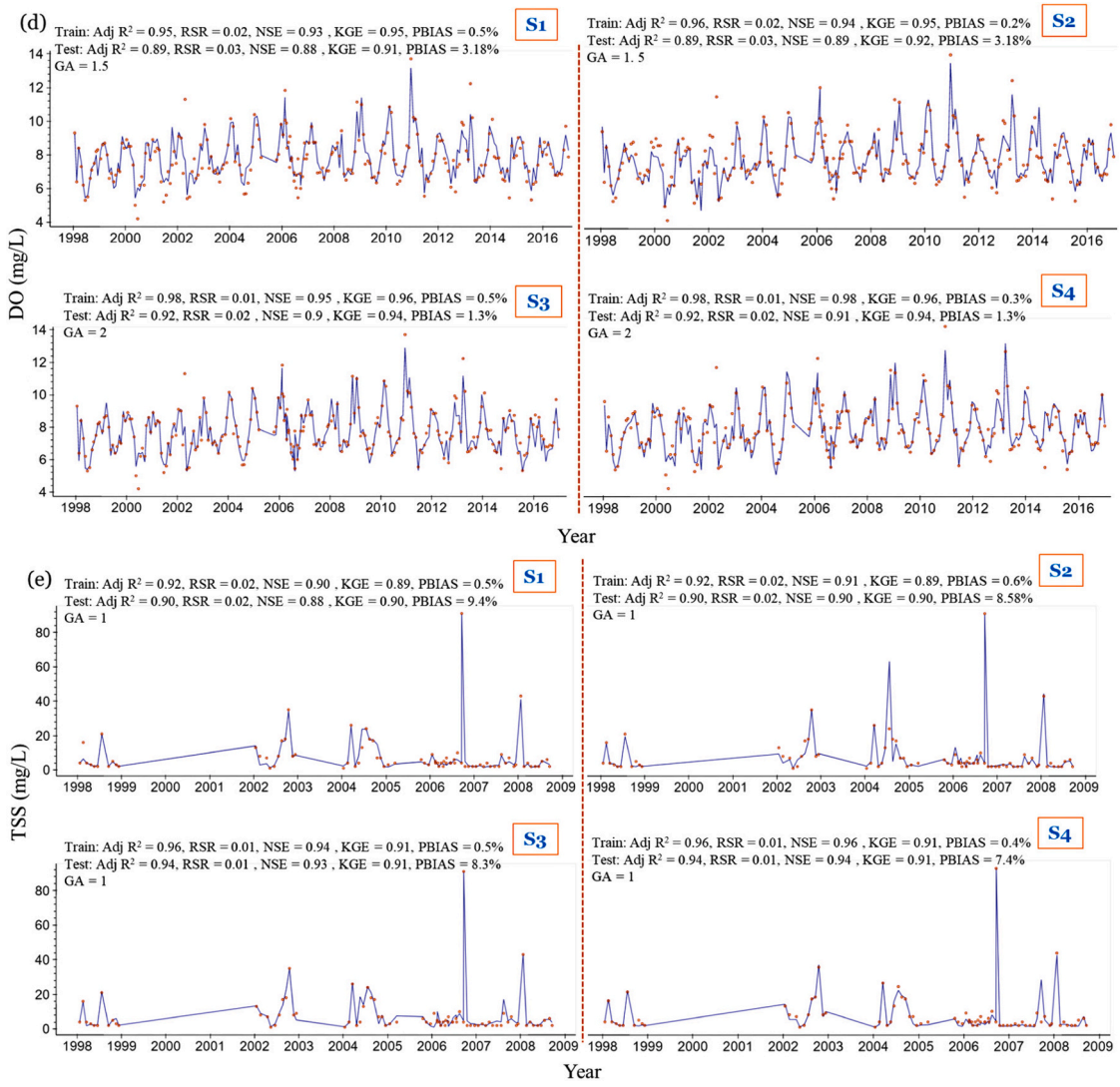


Fig. 7. (*continued*).

4.3. Importance of additional water quality constituents and antecedent conditions

Although the performance of each ML model somehow improved with the addition of antecedent conditions and water quality features, overall, the performance without their addition was satisfactory for all the constituents, except for FC. The model performance improved marginally (4.2% for FC, 2.2% for TP, and 1.2% for TN) with the inclusion of antecedent conditions in Scenario S2, consistent with previous studies that emphasized the association of water quality constituents like bacteria with prior rainstorm events due to sediment resuspension in the water column (Abimbola et al., 2020; Farnham and Lall, 2015; Motamarri and Boccelli, 2012). Furthermore, the inclusion of water quality data in Scenario S3 significantly improved the prediction accuracy (21.1% for FC, 17% for TP, and 14.8% for TN), as observed in other water quality studies (Abimbola et al., 2021; Park and Engel, 2015, 2014). Horowitz et al. (2001) also stated that limited water quality data could lead to imprecise waterbody pollutant predictions, further alluding to the importance of water quality data. In this study, when additional water quality constituents were included in the ML predictions (Scenario S3), there was a 24% increase in Adj R^2, 23% decrease in RSR, and 42% decrease in PBIAS, respectively, compared to Scenario S1 (Fig. 6). Similarly, when antecedent conditions were considered alongside additional water quality constituents (Scenario S4), there was an additional 6.9% increase in Adj R^2, 4% decrease in RSR, and 6.9% decrease in PBIAS, respectively, from Scenario S3 to S4 (Fig. 6). Compared to Scenario S4, R^2 and RSR only improved by 1.8% and 2.9% when antecedent conditions were considered, and no additional water quality constituents were included (Scenario S2). Similarly, higher peaks in the FC time series were better captured in Scenarios S3 and S4, as shown in Fig. 7a. These findings suggest that the importance of antecedent conditions becomes more pronounced when used in conjunction with additional water quality constituents. Across all the four modeling scenarios, the models were found to overestimate FC concentration values <800 CFU/100 mL and underestimate concentrations >1360 CFU/100 mL. On average, there was an underestimation bias across all the modeling scenarios due to the underestimation of high FC concentrations, which was significantly reduced when additional water quality constituents and antecedent conditions were accounted for in Scenarios S3 and S4 (Fig. 7a). Underestimation of large FC concentrations could be due to the unbalanced observation dataset since concentrations >1320 CFU/100 mL cover only 29% of the entire observations. Nevertheless, GA fluctuated across the modeling scenarios without a clear pattern suggesting that prediction generalization does not necessarily improve with the inclusion of additional water quality constituents alone but through the combination of additional water quality constituents with antecedent conditions.

Similarly, model performance for TN and TP (Fig. 6, Fig. 7b, and

Table 3
Seasonal concentrations of the observed water quality constituents.

Season	FC (CFU/100 mL)	TN (mg/L)	TP (mg/L)	DO (mg/L)	TSS (mg/L)
Wet	1603.16	1.48	0.31	6.76	8.58
Dry	2003.03	1.37	0.19	8.71	6.93

Fecal coliform (FC) total nitrogen (TN); total phosphorus (TP); dissolved oxygen (DO); and total suspended solids (TSS).

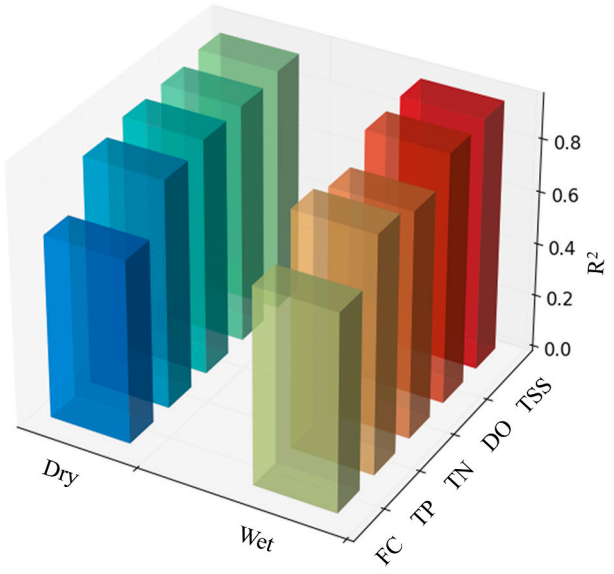


Fig. 8. Comparison of prediction performances of the best machine learning model—Artificial Neural Network (ANN), for all target water quality constituents, in the best modeling scenario (S4) during the wet and dry seasons. According to R^2, the prediction was more satisfactory for dry seasons for all the water quality constituents except TN. FC: Fecal coliform; TN: Total nitrogen; TP: Total phosphorus; DO: Dissolved oxygen; and TSS: Total suspended solids.

Fig. 7c) showed improvements in Scenarios S3 and S4, with the introduction of antecedent conditions and additional water quality constituents, although to a lesser degree compared to FC, suggesting that nutrients are less reliant on these factors than FC. In Scenario S3 (Fig. 4c and Fig. 7c), TN was less reliant on including additional water quality constituents than TP (Fig. 4b and Fig. 7b). The trend from Scenario S1 to S4 for TN showed that PBIAS did not significantly decrease with the introduction of additional water quality constituents. TN predictions had a slightly larger bias than TP, DO, and TSS in all the modeling scenarios' training phases, while FC had the largest PBIAS (~7.3%) in the training phase (Fig. 4c, Fig. 4d, and Fig. 7e). All evaluation criteria suggested that additional water quality constituents and antecedent conditions significantly improved the model performance for FC, TP, and TN; however, their influence in predicting DO and TSS was small (< 2% and < 1% improvement in R^2 for DO and TSS, respectively). Thus, using only publicly accessible datasets was sufficient to predict DO and TSS.

4.4. Seasonality of the water quality predictions

Predictive performances of water quality constituents were evaluated for wet (May to October) and dry (November to April) seasons across all the modeling scenarios. Table 3 shows that the average seasonal concentrations (1998–2016) were larger for FC and DO during the dry season but smaller for TP, TN, and TSS in this season. Almost identical model performances of all the water quality constituents during the wet and dry seasons suggested that the season-dependency of our predictions was minimal (<4% difference). Nevertheless, seasonal model performances showed that FC and TN had greater seasonal disparity compared to other target water quality constituents (Fig. 8). Considering the fit metrics in scenarios S1- S4, TN, and DO were slightly better predicted during the wet season, while FC, TP, and TSS were slightly better predicted during the dry season. Also, there was a clear seasonal differentiation for the nutrients. For TP, model performance was more profound in the dry season for three scenarios (S2, S3, and S4), while superior performances were found for the wet season for TN. This suggests that TN is greatly affected by season regardless of the modeling scenario. The effects of meteorologic features (rainfall, CR5, air temperature, and relative humidity) were more substantial during the dry season (i.e., leading to a better model performance). However, because there were no apparent differences in the seasonal importance of meteorological features for TP, the superior performance during the wet season cannot be directly attributed to meteorologic features. The superior predictive performance of all the water quality constituents in Scenario S1 can be attributed to more profound importance of rainfall in this scenario. From Scenario S1 to S4, the importance of rainfall dampens with the subsequent introduction of antecedent conditions and water quality constituents, and more prominently during the wet season for Scenario S4. This can explain why prediction performance for all target constituents, excluding TN, was superior during the dry season for Scenario S4. Similarly, although FC's concentration was higher during the dry season, there was more underestimation of large FC concentrations during the wet season; this provides an additional possible explanation for a poorer prediction performance during the dry season. Also, TSS and turbidity were notably prominent for improving FC, TP, and TSS predictions (Scenarios S3 and S4), highlighting the role of sediment transport during the dry season. This suggests that the effect of short intermittent rainfall events during the wet season are stronger drivers of these water quality constituents than more frequent and intense ones during the dry season.

4.5. Uncertainty quantification

Predictive uncertainties of the best ML models for each target constituent are presented in Fig. 9. The findings showed that FC and DO predictions were the most and least uncertain, respectively. Aside from DO, the predictions of water quality constituents were least uncertain for Scenario S4. Incorporating antecedent conditions and additional water quality constituents (Scenarios S2-S4) reduced the predictive uncertainty around predictions of all water quality constituents. This reduction was most significant for FC and TSS (Fig. 9a), corroborating the earlier evidence of accounting for water quality processes like bacteria resuspension that can be explained by water quality constituents like turbidity and TSS. Although TP predictions were more uncertain than TN, TN had a larger PBIAS on average (8.6% for TN, 5.5% for TP). Nevertheless, TP experienced greater improvements in PBIAS from Scenario S1 to S4 alongside FC and TSS (Fig. 6). Also, while DO and TSS had outstanding performances (Adj $R^2 > 0.9$; Fig. 9e and Fig. 9f), predictive uncertainties were larger for TSS. This can be attributed to the relatively small observation dataset of TSS (about one-third of the other water quality constituents). The uncertainties can be attributed to learning/optimization algorithm biases, water quality measurements, the approximation of features across time and space (when the data were not available), and inadequate features to capture specific watershed processes such as atmospheric deposition, nitrification/denitrification, among others (Mallya et al., 2020).

Since the uncertainty reported in this study depended on the existing data in our case study, the reliability of extrapolating the findings to a similar study area (low-lying with similar land cover distribution) depends on knowledge of watershed processes and data quality. However, our presented modeling framework is generic and applicable to modeling in-stream water quality at the watershed scale.

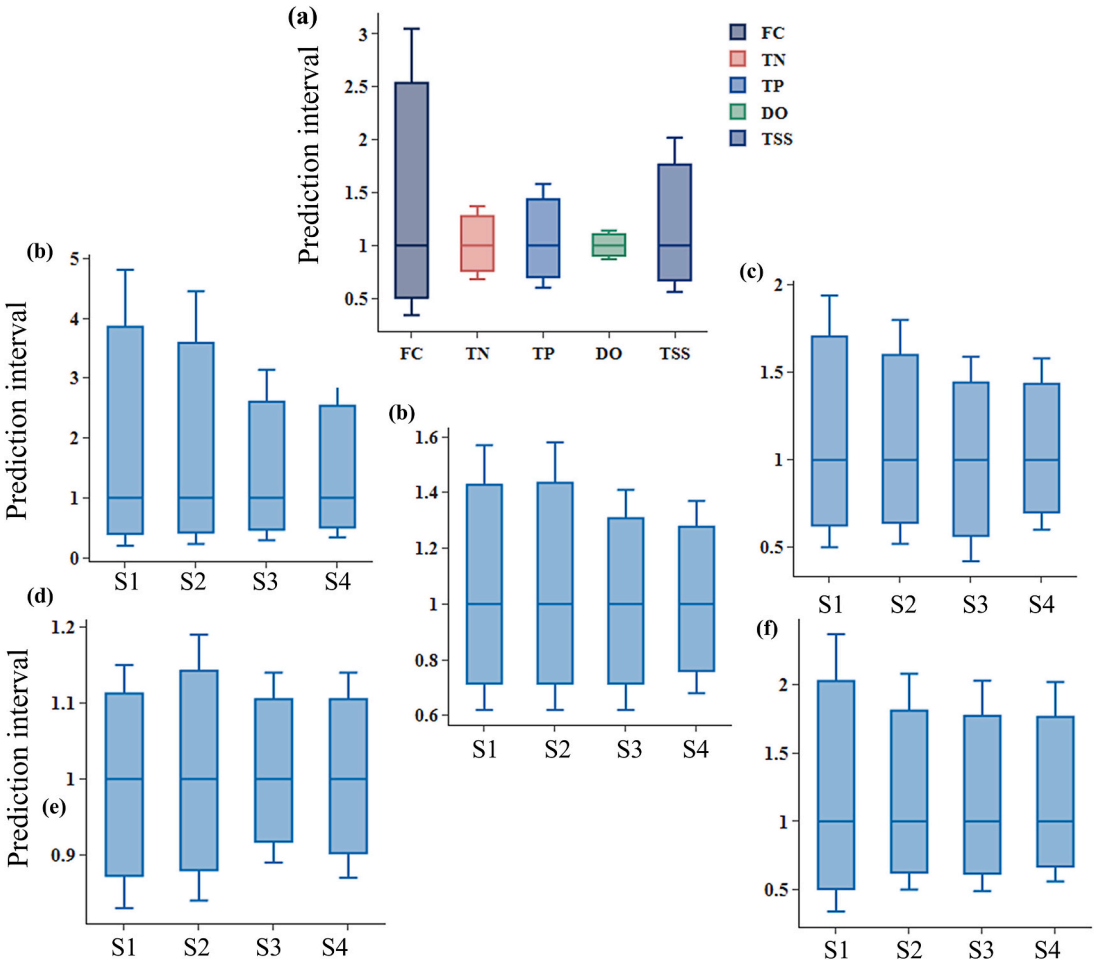


Fig. 9. Comparison of the ML models' uncertainty (prediction interval around hypothetical value '1'), of *'best model scenario (S4)'* (a) and *'all model scenarios (S1, S2, S3, and S4)'* of each target water quality constituents: fecal coliform (FC) (b), total nitrogen (TN) (c), total phosphorus (TP), dissolved oxygen (DO) (d) and total suspended solids (TSS) (e). Overall, FC and DO predictions had the largest and smallest uncertainties, respectively; meanwhile, the model uncertainty was always smallest in scenario S4 except for DO.

5. Summary and conclusions

This study demonstrated a ML-based framework to predict five in-stream water quality constituents—TP, TN, TSS, and DO—at the watershed scale. The framework used water quality drivers (meteorologic, hydrologic, geomorphologic, land cover, and pedologic), which represent pertinent physical, chemical, and biological processes, and are publicly available throughout the US alongside antecedent conditions and other water quality constituents that explain physical processes (pH and turbidity). We evaluated the performance of five ML algorithms—SVM, RF, XGB, RF-XGB, and ANN—using various fit metrics (Adj R^2, NSE, PBIAS, KGE, and GA). Explanatory variables representing water quality processes were identified using tree-based SHAP, ReliefF ranking, and spearman rank-order correlation. Feature analyses revealed that water quality constituents (TSS and turbidity) were the most important drivers of FC, TP, and TSS, while antecedent conditions (CR5) and meteorological factors like rainfall and air temperature were the most important for TN and DO's predictions. This finding, in addition to the model performance, generally revealed that:

1) Though including additional water quality drivers improved overall model performance for all target constituents, TP, TN, DO, and TSS could still be predicted satisfactorily using only publicly available datasets (Nash-Sutcliffe efficiency [NSE] $>$ 0.75 and percent bias [PBIAS] $<$ 10%), whereas FC could not (NSE $<$ 0.49 and PBIAS $>$25%).
2) Water quality data and antecedent conditions are influential in improving the predictive performance, capturing high concentrations, and in reducing predictive uncertainties for all target water quality constituents (FC, TP, TN, DO, and TSS). Despite these advantages, they were generally not necessary in predicting TP, TN, DO, and TSS (Adj $R^2 > 0.71$, NSE > 0.69, KGE > 0.72, and PBIAS $<$10%). In contrast, they are a necessity to achieve satisfactorily prediction for FC (Adj $R^2 < 0.68$, NSE < 0.68, KGE < 0.69, and PBIAS $<$14.3%) and even better when used in conjunction with antecedent conditions (Adj $R^2 < 0.72$, NSE < 0.7, KGE < 0.7, and PBIAS $<$13.4%).

   The most remarkable model improvement was observed for FC, followed by TP, TN, DO, and TSS based on Adj R^2 and NSE, whereas with regards to PBIAS, the most significant improvement was observed for FC, TP, and TSS.
3) Prediction uncertainties decreased for all target constituents and most prominent for FC and TSS, and the smallest predictive uncertainty was found for DO, followed by TN, TP, TSS, and FC. This showed that FC and DO predictions were the most and least uncertain, respectively.
4) Although there were disparities in the prediction performances of target water quality constituents during the wet and dry seasons, these differences were only marginal ($<$4% difference in the fit metrics) and more pronounced for FC and TN. Also, FC, TP, and TSS had better predictions during the dry season, while superior performances were obtained for TN and DO during the wet season. Thus,

short, sporadic dry season rainfall was more important in predicting FC, TP, and TSS, whereas longer and more frequent rainfall dominated in the predictions of TN and DO.

5) While all the examined ML algorithms performed adequately for all the target water quality constituents, RF-XGB and ANN produced more accurate and generalizable outputs.

This study shed insights into important water quality drivers and pertinent processes, using ML algorithms for predicting in-stream water quality, seasonality of these models, and the predictive uncertainties. The models can also serve as an alternative tool when process-based models cannot be implemented. This, in turn, can support water quality restoration projects like TMDLs in the U.S. and, more broadly, water quality restoration efforts. Future research should focus on evaluating the prediction performance of other ML algorithms like recurrent and attention-based neural networks, as well as autoregressive algorithms for water quality predictions. Similarly, evaluating other water quality constituents at different spatial scales and examining other watersheds with different spatiotemporal characteristics (e.g., tidally influenced and predominantly urban or agricultural) are potential research areas that needs to be explored.

CRediT authorship contribution statement

Itunu C. Adedeji: Conceptualization, Software, Methodology, Visualization, Data curation, Writing – original draft, Formal analysis, Investigation. **Ebrahim Ahmadisharaf:** Conceptualization, Methodology, Validation, Writing – original draft, Supervision. **Yanshuo Sun:** Methodology, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

This research was partially supported by the US National Science Foundation award numbers 2100745 and 2055347. The first author was partially supported by American Association of University Women's International Fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jconhyd.2022.104078.

References

Abba, S.I., Pham, Q.B., Saini, G., Linh, N.T.T., Ahmed, A.N., Mohajane, M., Khaledian, M., Abdulkadir, R.A., Bach, Q.V., 2020. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environ. Sci. Pollut. Res. 27 (33), 41524–41539. https://doi.org/10.1007/s11356-020-09689-x.

Abbas, A., Baek, S., Silvera, N., Soulileuth, B., Pachepsky, Y., Ribolzi, O., Boithias, L., Cho, K.H., 2021. In-stream *Escherichia Coli* Modeling Using high-temporal-resolution data with deep learning and process-based models. Hydrol. Earth Syst. Sci. Discuss. 1–55 https://doi.org/10.5194/hess-2021-98.

Abimbola, O.P., Mittelstet, A.R., Messer, T.L., Berry, E.D., Bartelt-Hunt, S.L., Hansen, S. P., 2020. Predicting *Escherichia coli* loads in cascading dams with machine learning: an integration of hydrometeorology, animal density and grazing pattern. Sci. Total Environ. 722, 137894 https://doi.org/10.1016/j.scitotenv.2020.137894.

Abimbola, O.P., Mittelstet, A., Messer, T., Berry, E., van Griensven, A., 2021. Modeling and prioritizing interventions using pollution hotspots for reducing nutrients, atrazine and e. Coli concentrations in a watershed. Sustainability (Switzerland) 13 (1), 1–22. https://doi.org/10.3390/su13010103.

Adams, H.D., Park Williams, A., Xu, C., Rauscher, S.A., Jiang, X., McDowell, N.G., 2013. Empirical and process-based approaches to climate-induced forest mortality models. Front. Plant Sci. 4 (NOV), 1–5. https://doi.org/10.3389/fpls.2013.00438.

Ahmadisharaf, E., Benham, B.L., 2020. Risk-based decision making to evaluate pollutant reduction scenarios. Sci. Total Environ. 702 (5), 719–727. https://doi.org/10.1016/j.scitotenv.2019.135022.

Ahmadisharaf, E., Camacho, R.A., Zhang, H.X., Hantush, M.M., Mohamoud, Y.M., 2019. Calibration and validation of watershed models and advances in uncertainty analysis in TMDL studies. J. Hydrol. Eng. 24 (7), 03119001 https://doi.org/10.1061/(asce)he.1943-5584.0001794.

Ahmadisharaf, E., Camacho, R.A., Zhang, H.X., Hantush, M.M., Mohamoud, M.Y., 2022. Model calibration and validation. In: Total Maximum Daily Load Development and Implementation: Models, Methods, and Resources. ASCE, Reston, VA, pp. 215–269.

Ahmadisharaf, E, Kalyanapu J, A, Thames A, B, Lillywhite, J, 2016. A probabilistic framework for comparison of dam breach parameters and outflow hydrograph generated by different empirical prediction methods. Environ. Model. Software 86, 248–263. https://doi.org/10.1016/j.envsoft.2016.09.022.

Alnahit, A.O., Mishra, A.K., Khan, A.A., 2022. Stream water quality prediction using boosted regression tree and random forest models. Stoch. Env. Res. Risk A. 4 https://doi.org/10.1007/s00477-021-02152-4.

Alqahtani, A., Shah, M.I., Aldrees, A., Javed, M.F., 2022. Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality. Sustainability (Switzerland) 14 (3). https://doi.org/10.3390/su14031183.

Al-Sulttani, A.O., Al-Mukhtar, M., Roomi, A.B., Farooque, A.A., Khedher, K.M., Yaseen, Z.M., 2021. Proposition of new ensemble data-intelligence models for surface water quality prediction. IEEE Access 9, 108527–108541. https://doi.org/10.1109/ACCESS.2021.3100490.

Banadkooki, F.B., Ehteram, M., Panahi, F., Sammen, Sh., Othman, F.B., EL-Shafie, A., 2020. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. J. Hydrol. 587 (February), 124989 https://doi.org/10.1016/j.jhydrol.2020.124989.

Beven, K., 2018. Environmental Modelling: An Uncertain Future? CRC Press.

Bicknell, B.R., Imhoff, J.C., Kittle Jr., K.L., Donigian Jr., A.S., 2005. Hydrological Simulation Program—Fortran (HSPF) user's Manual for Release 12.2. USEPA, Athens, GA.

Bilali, E.L., Taleb, A., Bahlaoui, M.A., Brouziyne, Y., 2021. An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset. J. Hydrol. 599 (May), 126510 https://doi.org/10.1016/j.jhydrol.2021.126510.

Borah, D.K., Yagow, G., Saleh, A., Barnes, P.L., Rosenthal, W., Krug, E.C., Hauck, L.M., 2006. Sediment and nutrient modeling for Tmdl development and implementation. Trans. ASABE 49 (4), 967–986. https://doi.org/10.13031/2013.21742.

Borah, D.K., Ahmadisharaf, E., Padmanabhan, G., Imen, S., Mohamoud, Y.M., 2019. Watershed models for development and implementation of total maximum daily loads. J. Hydrol. Eng. 24 (1).

Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020. A review of the artificial neural network models for water quality prediction. Appl. Sci. (Switzerland) 10 (17). https://doi.org/10.3390/app10175776.

Cho, K.H., Pachepsky, Y.A., Oliver, D.M., Muirhead, R.W., Park, Y., Quilliam, R.S., Shelton, D.R., 2016. Modeling fate and transport of fecally-derived microorganisms at the watershed scale: state of the science and future opportunities. Water Res. 100, 38–56. https://doi.org/10.1016/j.watres.2016.04.064.

Copeland, C., 2012. Clean Water Act and Pollutant Total Maximum Daily Loads: CRS Report for Congress, pp. 1–18. http://www.crs.gov.

Dada, A.C., Hamilton, D.P., 2016. Predictive models for determination of *E. coli* concentrations at inland recreational beaches. Water Air Soil Pollut. 227 (9) https://doi.org/10.1007/s11270-016-3033-6.

David, M.M., Haggard, B.E., 2011. Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River watershed, Arkansas and Oklahoma, USA. Water Air Soil Pollut. 215 (1–4), 525–547. https://doi.org/10.1007/s11270-010-0497-7.

Duan, W., Takara, K., He, B., Luo, P., Nover, D., Yamashiki, Y., 2013. Spatial and temporal trends in estimates of nutrient and suspended sediment loads in the Ishikari River, Japan, 1985 to 2010. Sci. Total Environ. 461–462, 499–508. https://doi.org/10.1016/j.scitotenv.2013.05.022.

Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A., Bahamonde, A., 2021. Scalable feature selection using ReliefF aided by locality-sensitive hashing. Int. J. Intell. Syst. 36 (11), 6161–6179. https://doi.org/10.1002/int.22546.

Elshorbagy, A., Teegavarapu, R.S.V., Ormsbee, L., 2005. Framework for assessment of relative pollutant loads in streams with limited data. Water Int. 30 (4), 477–486. https://doi.org/10.1080/02508060508691892.

Farnham, D.J., Lall, U., 2015. Predictive statistical models linking antecedent meteorological conditions and waterway bacterial contamination in urban waterways. Water Res. 76, 143–159. https://doi.org/10.1016/j.watres.2015.02.040.

Fluke, J., González-Pinzón, R., Thomson, B., 2019. Riverbed sediments control the spatiotemporal variability of *E. coli* in a highly managed, Arid River. Front. Water 1 (November). https://doi.org/10.3389/frwa.2019.00004.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Holcomb, D.A., Messier, K.P., Serre, M.L., Rowny, J.G., Stewart, J.R., 2018. Geostatistical prediction of microbial water quality throughout a stream network using

meteorology, land cover, and spatiotemporal autocorrelation. Environ. Sci. Technol. 52 (14), 7775–7784. https://doi.org/10.1021/acs.est.8b01178.
Horowitz J, A, Elrick A, K, Smith J, J, 2001. Estimating suspended sediment and trace element fluxes in large river basins. Hydrol. Process. 15 (7), 1107–1132. https://doi.org/10.1002/hyp.206.
Johnson, S.L., Maidment, D.R., Kirisits, M.J., 2013. TMDL balance: A model for coastal water pollutant loadings. J. Am. Water Resour. Assoc. 49 (4), 838–850. https://doi.org/10.1111/jawr.12044.
Kao, I.F., Zhou, Y., Chang, L.C., Chang, F.J., 2020. Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting. J. Hydrol. 583 (November 2019), 124631 https://doi.org/10.1016/j.jhydrol.2020.124631.
Kelderman, K., Phillips, A., Pelton, T., Schaeffer, E., Falcon, P., 2022. The Clean Water Act at 50 (Paper Knowledge. Toward a Media History of Documents).
Khatri, N., Khatri, K.K., Sharma, A., 2020. Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. J. Water Process Eng. 37 (March), 101477 https://doi.org/10.1016/j.jwpe.2020.101477.
Kononenko, I, 1994. Estimating attributes: Analysis and extensions of RELIEF. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics 171–182. https://doi.org/10.1007/3-540-57868-4_57.
Kuhn, M., Johnson, K., 2019. Feature engineering and selection: a practical approach for predictive models. Feature Eng. Selection: A Practical Approach Predictive Models 1–297. https://doi.org/10.1201/9781315108230.
Lamontagne, J.R., Barber, C.A., Vogel, R.M., 2020. Improved estimators of model performance efficiency for skewed hydrologic data. Water Resour. Res. 56 (9), 1–25. https://doi.org/10.1029/2020WR027101.
Lee J, C, Hirsch M, R, Schwarz E, G, Holtschlag J, D, Preston D, S, Crawford G, C, Vecchia V, A, 2016. An evaluation of methods for estimating decadal stream loads. J. Hydrol. 542, 185–203. https://doi.org/10.1016/j.jhydrol.2016.08.059.
Li, L., Qiao, J., Yu, G., Wang, L., Li, H.Y., Liao, C., Zhu, Z., 2022. Interpretable tree-based ensemble model for predicting beach water quality. Water Res. 211 (January), 118078 https://doi.org/10.1016/j.watres.2022.118078.
Mallya, G., Gupta, A., Hantush, M.M., Govindaraju, R.S., 2020. Uncertainty quantification in reconstruction of sparse water quality time series: implications for watershed health and risk-based TMDL assessment. Environ. Model. Softw. 131 (May), 104735 https://doi.org/10.1016/j.envsoft.2020.104735.
Mishra, A., Ahmadisharaf, E., Benham, B.L., Wolfe, M.L., Leman, S.C., Gallagher, D.L., Reckhow, K.H., Smith, E.P., 2018. Generalized likelihood uncertainty estimation and Markov chain Monte Carlo simulation to prioritize TMDL pollutant allocations. J. Hydrol. Eng. 23 (12), 05018025 https://doi.org/10.1061/(asce)he.1943-5584.0001720.
Mishra, A., Ahmadisharaf, E., Benham, B.L., Gallagher, D.L., Reckhow, K.H., Smith, E.P., 2019. Two-phase Monte Carlo simulation for partitioning the effects of epistemic and aleatory uncertainty in TMDL modeling. J. Hydrol. Eng. 24 (1), 04018058 https://doi.org/10.1061/(asce)he.1943-5584.0001731.
Moriasi, D.N., Guzman, J.A., Steiner, J.L., Starks, P.J., Garbrecht, J.D., 2014. Seasonal sediment and nutrient transport patterns. J. Environ. Qual. 43 (4), 1334–1344. https://doi.org/10.2134/jeq2013.11.0478.
Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: performance measures and evaluation criteria. Trans. ASABE 58 (6), 1763–1785. https://doi.org/10.13031/trans.58.10715.
Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res. 46 (14), 4508–4520. https://doi.org/10.1016/j.watres.2012.05.023.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Neitsch, S., Arnold, J., Kiniry, J., Williams, J., Lantagne, D.S., Blount, B.C., Cardinali, F., Quick, R., Whitehead, P.G., Leckie, H., Rankinen, K., Butterfield, D., Futter, M.N., Bussi, G., Bicknell, B.R., Imhoff, J.C., Kittle, J.L., Donigian, A.S., Johanson, R.C., 2011. Soil & water assessment tool theoretical documentation version 2009. Texas Water Res. Inst. 572 (June), 1–647. https://doi.org/10.1016/j.scitotenv.2015.11.063.
Park S, Y, Engel A, B, 2015. Analysis for regression model behavior by sampling strategy for annual pollutant load estimation. J. Environ. Qual. 44 (6), 1843–1851. https://doi.org/10.2134/jeq2015.03.0137.
Park, Y.S., Engel, B.A., 2014. Use of pollutant load regression models with various sampling frequencies for annual load estimation. Water (Switzerland) 6 (6), 1685–1697. https://doi.org/10.3390/w6061685.
Runkel, R.L., Crawford, C.G., Cohn, T.A., 2004. Load Estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers. In: Techniques and Methods. U.S. Geological Survey, vol. 4. U.S. Department of the Interior, p. 69.
Sakizadeh, M., 2016. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Sys. Environ. 2 (1), 52–62. https://doi.org/10.1007/s40808-015-0063-9.
Schwarz, G.E., Hoos, A.B., Alexander, R.B., Smith, R.A., 2006. The SPARROW Surface Water-Quality Model: Theory, Application and User Documentation: U.S. Geological Survey Techniques and Methods Book 6, Section B, Chapter 3, p. 248.
Sigleo, A., Frick, W., 2003. Seasonal variations in river flow and nutrient concentrations in a northwestern USA watershed. In: … on Research in the Watersheds. US …, *Figure 1*. In: http://www.tucson.ars.ag.gov/icrw/Proceedings/Sigleo.pdf.
Stocker, M.D., Pachepsky, Y.A., Hill, R.L., 2022. Prediction of *E. coli* concentrations in agricultural pond waters: application and comparison of machine learning algorithms. Front. Artif. Intell. 4 (January), 1–12. https://doi.org/10.3389/frai.2021.768650.
Tripathi, S., Govindaraju, R.S., 2007. On selection of kernel parametes in relevance vector machines for hydrologic applications. Stoch. Env. Res. Risk A. 21 (6), 747–764. https://doi.org/10.1007/s00477-006-0087-9.
Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., Moore, J.H., 2018. Relief-based feature selection: introduction and review. J. Biomed. Inform. 85, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014.
USEPA, 2001. National Costs to Implement TMDLs (Draft Report): Support Document 2. August, p. 176.
Wahl, T.L., 2004. Uncertainty of predictions of embankment dam breach parameters. J. Hydraul. Eng. 130 (5), 389–397. https://doi.org/10.1061/(asce)0733-9429(2004)130:5(389).
Wang, R., Kim, J.H., Li, M.H., 2021. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Sci. Total Environ. 761, 144057 https://doi.org/10.1016/j.scitotenv.2020.144057.
Yu, X., Shen, J., Du, J., 2021. An inverse approach to estimate bacterial loading into an estuary by using field observations and residence time. Mar. Environ. Res. 166 (October 2020), 105263 https://doi.org/10.1016/j.marenvres.2021.105263.
Zhang, H.X., Quinn, N.W.T., 2019. Simple models and analytical procedures for total maximum daily load assessment. J. Hydrol. Eng. 24 (2), 02518002 https://doi.org/10.1061/(asce)he.1943-5584.0001736.
Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: a review. J. Hydrol. 598 (April), 126266 https://doi.org/10.1016/j.jhydrol.2021.126266.