**Prediction of the Concentration of Dissolved Oxygen in Running Water by Employing A Random Forest Machine Learning Technique**

Mohammad Hafez Ahmed

Department of Civil and Environmental Engineering, West Virginia University, PO BOX 6103, Morgantown, WV 26506-6103, USA
Correspondence: mha0015@mix.wvu.edu; Tel: +1-304-216-9519

**Abstract:** Dissolved oxygen (DO) is a key indicator in the study of the ecological health of rivers. Modeling DO is a major challenge due to complex interactions among various process components of it. Considering the vital importance of it in water bodies, the accurate prediction of DO is a critical issue in ecosystem management. Given the intricacy of the current process-based water quality models, a data-driven model could be an effective alternative tool. In this study, a random forest machine learning technique is employed to predict the DO level by identifying its major drivers. Time-series of half-hourly water quality data, spanning from 2007 to 2019, for the South Branch Potomac River near Springfield, WV, are obtained from the United States Geological Survey database. Key drivers are identified, and models are formulated for different scenarios of input variables. The model is calibrated for each input scenario using 80% of the data. Water temperature and pH are found to be the most influential predictors of DO. However, satisfactory model performance is achieved by considering water temperature, pH, and specific conductance as input variables. The model validation is made by predicting DO concentrations for the remaining 20% of the data. The comparison with the traditional multiple linear regression method shows that the random forest model performs significantly better. The study insights are, therefore, expected to be useful to estimate stream/river DO levels at various sites with a minimum number of predictors and help build a sturdy framework for ecosystem health management across an environmental gradient.

**Keywords:** dissolved oxygen; rivers; random forest; multiple linear regression; water quality; ecosystem health

1. Introduction

Surface water quality is deteriorating globally due to the high level of pollutant loads [1]. Overall, 40 percent of the rivers have become polluted to varying extents [2]. Pollution has become a major concern due to its increasing trend. It is also posing serious threats to ecological integrity that includes degraded health of aquatic water bodies (i.e., habitat instability of aquatic life), aesthetical nuisance (i.e., algal bloom), hypoxia (i.e., dissolved oxygen < 2.0 mg/L), and so on [3]. Therefore, maintaining and protecting a healthy stream is a priority for a sustainable ecosystem. Dissolved oxygen (DO) is considered one of the most important key indicators that is used to evaluate the biological health of rivers [4,5]. It is highly desirable to maintain a minimum DO level (e.g., ~5.0 mg/L) for the survival of diverse aquatic life [6,7]. Understanding DO dynamics and its effective predictions, therefore, are of critical importance for the design and operation of a sustainable ecosystem.

The main sources of in-stream DO include atmospheric air-water interaction (i.e., reaeration) and aquatic plant photosynthesis [4]. Denitrification and external inputs also act as DO sources. DO sinks include decomposition of carbonaceous organic matter, nitrification, and aquatic respirations [8]. The identification and quantification of these processes are challenging for researchers and water resource managers. Further, the coupling of natural factors and human-induced influences (e.g., agricultural activities, urban sprawling) complicate these interactions up by one more level [9].
Although DO is primarily water temperature-dependent, the underlying factors are subject to vary due to hydro-meteorological alterations [10] and the intensity of the biochemical processes. Therefore, it is highly desirable to develop a DO model for each major water body considering its ecological importance [11] – leading to an optimized water quality monitoring system throughout a time horizon.

Proper monitoring and quantification of the DO level have become a global concern for water quality managers and policy decision-makers. However, precise DO levels prediction is a formidable challenge because of the interactions among different process components. Realization of its utmost importance led to the inclusion of DO in water quality modeling. At the commencement of the water quality model development, a simple DO model was established by Streeter-Phelps in 1925 [12]. After that much effort has been devoted to building various water quality models, where DO is considered as an inherent component. Readers are referred to elsewhere in order to get a detailed overview of the publicly available water quality models for DO simulations [13,14]. These models were mechanistically developed and are mainly the modified versions of the original Streeter-Phelps equation. Mechanistic models are based on the mass-balance theory [4], encompassing the descriptions of the internal mechanisms of a system and explain more about the system behavior [15].

While these process-based models have proven to be useful, they have their own limitations (e.g., subjective model assumptions and consent setting goals). Most of these models require a huge amount of input data for simulations and subsequent validations [13]. This is computationally expensive and at the same time financially unattractive. There is a further need for a certain level of expertise. Sometimes these required data sets are not always readily available; thus, limits the potential implementation of models in many instances. The drawbacks of these process-based models bring the necessity of data-driven models in practice; which is both relatively simple and cost-effective.

Data-driven model development is a long-practiced technique for researchers and model developers. It is easy to implement and deals only with the inputs and outputs of a physical phenomenon. Though numerous DO models in aquatic systems have been established, the proper implementation of a specific model is subject to change due to its highly nonlinear dynamics and lack of adequate data sets [11,16]. In numerous cases, statistical relationships cannot be established (i.e., poor statistical significance and poor model prediction). This is the important underlying motivation to predict DO by applying a data-driven (i.e., machine learning) method and assess its performance.

Extensive literature reviews suggest that predictive model development of stream DO by implementing data-driven techniques has acquired considerable traction across the scientific community of stream water quality. Artificial neural networks (ANN) have been applied in a myriad water quality modeling studies including DO predictions. For instance, DO was predicted using an ANN model in the Te-Chi reservoir, Taiwan [17]. A feed-forward neural network (FNN) was applied in predicting DO in the Gruza reservoir, Serbia [11]. Estimation and comparison of upstream DO were made by using a multi-layer perceptron neural network (MLPNN) and radial basis function neural network (RBFNN) at two USGS stations [18]. DO concentrations were predicted using a generalized regression neural network (GRNN) and compared with traditional multiple linear regression (MLR) DO model in the Upper Klamath River, Oregon, USA [19]. Zhang et al. [20] applied a combined Kernal Principal Component Analysis (kPCA) and Recurrent Neural Network (RNN) to forecast the trend of dissolved oxygen in the Burnett river, southern Queensland.

Versatile fuzzy logic and neural fuzzy models were developed in different studies to predict the DO level as well. For example, fuzzy logic modeling was used to predict the DO level in the Golden Horn [21]. Two adaptive neuro-fuzzy techniques were applied to model hourly DO concentrations at Klamath River, Oregon, USA [22]. DO levels were computed using a neuro-fuzzy technique at Lower Broad River, South Carolina, USA [23]. Chen and Liu [24] modeled DO in a reservoir in Taiwan using back propagation neural network (BPNN) and adaptive neural-based fuzzy inference system (ANFIS) approaches. Nemati et al. [25] adopted ANN-ANFIS models to estimate DO concentrations in the Tai Po River, Hong Kong. Also, the support vector machine (SVM) is progressively gaining a higher level of interest in predictive water quality model development. It has been extensively
applied in numerous water resources studies. For example, SVM was employed to predict DO by employing five water quality variables in the Terengganu River, Malaysia [26]. Further, SVM was applied to predict DO concentration in a hypoxic river in southeastern China [27]. In a very recent study, the least square SVM was applied to predict DO from four water quality variables at three different USGS stations [28].

The neural network was the most mentioned data-driven technique alongside other machine learning algorithms in DO prediction. Given the availability of different data-driven methods, random forests are somehow underused and often underestimated in water resources for ambiguous reasons, especially for water quality predictions. However, random forests were successfully implemented in several studies for hydrological prediction. For instance, Wang et al. [29] used a random forest to assess flood hazard risk in Dongjiang River Basin, China. A spring discharge was forecasted using a random forest in Umbra region, Italy [30]. Random forests to estimate the evapotranspiration were implemented in separate studies by Granata [31] and Granata et al. [32]. In a different study, groundwater potential map was produced, using a random forest by Naghibi et al. [33]. The random forest technique has also been routinely applied in a variety of industrial and scientific research areas. For example, protein-protein interactions were predicted using a random decision forest framework [34]. A random forest regression was employed in predicting soil surface texture in a semiarid region [35]. Further, in separate studies, random forest networks were adopted in agricultural production systems, i.e., prediction of crop yield [36] and in building energy prediction [37]. These examples underscore the wide applications of random forests technique across disciplines. In addition, the application of a random forest model can extract the most influential drivers of a response variable from myriad interacting variables which eventually can lead to building a parsimonious predictive model. This advantage further motivated to apply this method in the present study.

This study aims to identify the key drivers of DO in a river and assess the ability to perform of the random forest machine learning algorithm in predicting DO levels for various combinations of input variables as well as to select the best model. A certain part of data was randomly selected for model development (i.e., model calibration) and the remaining part for model testing (i.e., model validation). To the best of knowledge, this particular machine learning algorithm is quite new in predicting half-hourly DO concentrations. A brief discussion of this technique is given in the method section (Section 2.4).

2. Materials and Methods

2.1. Study area and data collection

Historical time-series of half-hourly DO data and associated water quality variables (see Figure A1 in Appendix A) were retrieved from the United States Geological Survey (USGS) publicly open nation’s stream water quality database [38]. South Branch Potomac River near Springfield, WV, was selected as the study area (USGS ID: 1608500, Latitude: 39.44704, Longitude: -78.6542) (Figure 1). The natural drainage area of this study site is approximately 3,784 km² and is located at an elevation of 171.2 m from the mean sea level. The contributing catchment area of this station is vegetation dominated (i.e., ~82%, mainly deciduous forest). This area is still free from urban-sprawling (i.e., developed area = ~3.5%) – indicating that the selected river is naturally influenced. However, a small agricultural activity persists (i.e., ~14%). Other land-use activities and/ presence of open water bodies are comparatively negligible. Different land-use types were estimated using recently published NLCD data in ArcGIS 10.6. The NLCD is a national land use/land cover data set of the conterminous USA and can easily be retrieved at the USGS website [39]. The temporal variation was conserved using data spanning from 2007 to 2019. This area is selected as a potential case study to evaluate the
impact of water quality drivers on the overall ecological health in term of dissolved oxygen given the river is less affected by urbanization/agricultural activities as well as long-term data availability.

Figure 1. The water quality monitoring station and corresponding watershed boundary.

Predictor variables include water temperature ($T_w$; °C), instantaneous discharge ($Q$; m$^3$/s), turbidity (FNU), gage height (GH; m), water pH, and specific conductance (SC; μS/m). DO has a direct dependence on water temperature [10]. It primarily regulates the physical and biogeochemical processes in an environmental system. For example, it impacts the chemical reaction and dissolution capacity of gases and solids in the water column [40,41]. Discharge represents the reaeration effect on stream DO because of turbulent diffusion via atmospheric exchange, where gage height is proportionally related to discharge [42]. Turbidity was used as a surrogate variable of total solids. The pH is a measure of the effective hydrogen-ion concentration that is typically controlled by the interacting chemical reactions. The specific conductance is also an indirect measure of salinity that reduces the dissolution capacity of DO [4]. Detailed statistical descriptions of the data set were reported in Table 1. The data matrix for the main analysis was formed based on data quality and scientific relevance. Data only approved for publication was considered, while provisional data subject to revision was discarded. Any missing data of an individual predictor along with other co-measured variables were removed in order to get a complete data matrix. This noise removal process limited the final data points to ~200 thousand.
Table 1. Statistics of water quality variables and correlation between DO with predictor variables.

| Variables | Mean value | Standard deviation | Minimum value | 25th percentile | 50th percentile | 75th percentile | Maximum value | r  |
|-----------|------------|--------------------|---------------|-----------------|-----------------|-----------------|---------------|----|
| $T_w$ (°C) | 14.78      | 9.02               | -0.1          | 6.2             | 14.7            | 23.5            | 33.6          | -0.90 |
| $Q$ (m$^3$/s)$\times$0.028 | 1575      | 2372               | 48.5          | 310             | 846             | 1830            | 33800         | 0.11  |
| SC (μS/m)$\times$0.01 | 217.91    | 47.64              | 105           | 180             | 212             | 257             | 329           | -0.38 |
| pH        | 8.17       | 0.39               | 6.0           | 7.9             | 8.1             | 9.4             | 9.8           | 0.07  |
| TUR (FNU) | 11.16      | 43.80              | 0             | 1.2             | 2.6             | 6.6             | 1840          | -0.06 |
| GH (m)$\times$0.305 | 2.96      | 1.80               | 0.85          | 1.82            | 2.51            | 3.44            | 19.4          | 0.16  |
| DO (mg/L) | 10.49      | 2.31               | 5.1           | 8.6             | 10.2            | 12.3            | 18.8          | 1.00  |

Note: $T_w$, $Q$, SC, pH, TUR, GH, and DO denotes, respectively, water temperature, discharge, specific conductance, water pH, turbidity, gage height, and dissolved oxygen. The symbol ‘r’ stands for correlation coefficient.

Data normality of each predictor variable was checked by employing the Kolmogorov-Smirnov test [43]. Data distributions were compared with normal distributions by setting the significance level at 5 percent. The estimated statistics ($p$ values < 0.05) rejected the null hypothesis that the distributions come from normal distributions. The non-normality of data warranted the application of data transformation that approximately conserves normality. Therefore, data sets were Box-Cox transformed [44]. The Box-Cox transformation technique transforms non-normally distributed data into a set of data that has an approximately normal distribution. It is a family of power distributions and transformed the original data as follows, Eq. (1a-1b):

\[
X = \left( Y^\alpha - 1 \right) / \alpha; \quad \text{if } \alpha \neq 0 \\
X = \ln(Y) \quad \text{[log base } e\text{]; if } \alpha = 0
\]

where $Y$ is the original data; $\alpha$ (alpha) is the transformation parameter that maximizes the Log-Likelihood function; and $X$ is the transformed value.

Special condition when $\alpha = 0$:

\[
X = \left( Y^\alpha - 1 \right) / \alpha = (e^{\alpha \ln(Y)} - 1) / \alpha \\
X = ((1 + \alpha \ln(Y)) + 1/2! (\alpha \ln(Y))^2 + 1/3! (\alpha \ln(Y))^3 + \ldots) - 1) / \alpha \\
X = \alpha \ln(Y) + 1/2! (\alpha \ln(Y))^2 + 1/3! (\alpha \ln(Y))^3 + \ldots \\
X = \ln(Y); \quad \text{if } \alpha = 0
\]

Since the Box-Cox method works on only positive data, the water temperature was converted into the Kelvin scale to avert negative temperature. Further, zero turbidity values were also replaced by a small fraction of positive quantity which is closer to zero. The alpha values for $T_w$, $Q$, SC, pH, turbidity, and gage height were computed, respectively, as 0.55, -0.012, 0.38, -1.33, 0.012, and 0.46.

Data standardization was also done using the Z-score method, Eq. (2), in order to bring different units under a common reference scale that reduced the scale effects in the estimation of the response variable. Employment of transformed data has also considerably ameliorated model performance.

\[
Z_i = (x_i - \bar{x}_m) / S_d
\]
where $Z_i = \text{the standardized value}$; $x_i = \text{the observed value}$; $x_m$ and $S_d$ are, respectively, the mean value and standard deviation of a predictor variable. The approximate normal distribution of each predictor variables was visually checked by histograms (Figure 2).

![Histograms of all predictor variables using boxcox and Z-score transformed values.](image)

**Figure 2.** Histograms of all predictor variables using boxcox and Z-score transformed values. $T_w$, $Q$, SC, pH, TUR, GH, represents, respectively, water temperature, discharge, specific conductance, water pH, turbidity, and gage height.

2.2 Correlation coefficients

The correlation matrix was formed by estimating the Pearson correlation coefficients. The matrix helps glean background information among the variables. The cell values of the correlation matrix represent the stochastic connections of nonlinear correspondences by measuring the strength and direction between two corresponding variables. The important information on the presence of possible multicollinearity (i.e., mutual correlations between variables) can also be visualized by looking at the cross-correlations. The correlation coefficient was reported at the 95% confidence intervals ($p$ value < 0.05). The correlation matrix was formed on the transformed domain by using Box-Cox transformed and standardized data.

2.3 Important variable selection

Mutually correlated predictor variables with the response variable possibly contributes bias in the model performance [45]. Further, multiple parameter sets might potentially lead to the equifinality in a model [46]. Meaning that it is possible to provide a good fit between modeled and observed data from a combination of predictors. It is therefore critical to select an optimal number of mechanistically meaningful predictor variables for data-driven model building [47]. It is a delicate process and the random removal of predictors might result in bias estimation. Therefore, before identifying the limited number of key predictors, the random forest model was fitted with all available variables and a large number of trees. Then a variable importance metrics (VIMs) was
determined that provides the variable importance score. The score of each variable was estimated using Eq. (3) adopted by Ai et al. [48].

\[
VIS_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} D_{Gkij}}{\sum_{k=1}^{m} \sum_{i=1}^{n} D_{Gkij}}
\]  

(3)

where \( m, n, \) and \( t \) are respectively the total number of attributes, decision trees, and nodes. \( D_{Gkij} \) is the Gini decrease value (see in Section 2.4) of the \( j \)th node in the \( i \)th tree that belongs to the \( k \)th attribute. \( VIS_k \) is the variable importance score of the \( k \)th attribute. This matrix therefore unravels the relative significance (i.e., predictive capability) of each predictor in modeling the behavior of the response variable [49].

However, it cannot provide the exact information regarding the absolute performance of individual predictor variables in predicting the model output. Despite this, it is still useful to select the important variable that demonstrates a significant impact on the model response. Using the top-ranked predictor (chosen based on VIMs), the first random forest was created, and the model performance was evaluated by computing model statistics. Repeating the process and by adding the next important variable the random forest was recreated and the corresponding model performance was reevaluated. It stopped at a point where adding variables did not considerably improve the overall model efficacy. With this procedure, the most influential predictors were selected and employed in the final model. This finalized model can be termed as an “optimal /reduced model” – containing a minimum set of variables that maximized the model accuracy in predicting the model output.

2.4 Random forests regression model

The random forest is a widely used supervised machine learning technique used for both regression and classification problems. As the target variable (i.e., model output) is continuous, the regression algorithm was applied. It aggregates numerous decision trees (usually user-defined) and obtains a summary table of the model output [50]. It takes ‘\( n \)'-number’ sub-training sample sets from the total training sample set. This randomized resampling is recursively done by employing the bootstrapping method with data replacement. In each case, decision trees are established for the different combinations of predictor variables. With each decision tree, the response of the model is estimated and stored. From the multitude of decision trees – which constructs a random forest – a mean response is estimated as the final model output. The selection of the number of decision trees is critical in random forest analysis. It can range from a few hundred to several thousand. An optimum number of trees can be determined using a k-fold cross-validation approach [51]. However, given the huge amount of data, the number of decision tress was kept 100 as default to avoid the computational nuisance.

The trained model is determined by minimizing prediction errors. Once the prediction error is optimally minimized, the model is considered as the best-trained model. The total error is minimized using Eq. (4) for each decision tree of each node.

\[
GINI(P) = 1 - \sum_{i=1}^{n} p_{i}^2
\]  

(4)
where GINI (P) is the impurity index of a particular node expressed as probability; \( p_i \) is the probability of an attribute in each node that belongs to a particular sub-sample set. At each point of data splitting, the probability is estimated – where the minimum GINI represents the best estimates as it contains the least impurity. The overall flow chart of the random forest regression is given in order to get a basic understanding of how the algorithm functions (Figure 3).

![Flow chart of random forest algorithm](image)

**Figure 3.** Flow chart of a random forest algorithm. Total ‘n’ sub-sample sets are randomized and produce n decision trees. Y denotes the model output.

A detailed description of how a decision tree is formed and subsequent random forest works can be found elsewhere [49,50]. All the mathematical and statistical calculations of the random forest algorithm were made in Python 3.7 ‘sklearn.ensemble’ module and Excel 2016 programming platform.

### 2.5 Multiple linear regression (MLR) model

In order to evaluate the performance of a random forest model, the traditional MLR model was employed as a reference. The MLR model is composed of linear combinations of multiple predictor variables and regression weights and was developed as follows, Eq. (4):

\[
y = c + \sum_{i=1}^{n} b_i x_i
\]  

(4)

Where \( y \) = response variable; \( c \) = y-intercept; \( i \) = number of predictor variables (1, 2, 3, … n); \( b_i \) = regression weights; and \( x_i \) = predictor variables.
2.6 Data partitioning

It is important to note that the model development using training data set and model testing using a completely new data set were made following the 80/20 data partition as a rule of thumb [51]. Total 80% of the filtered data (~160 thousand data points) was randomly selected in model calibration (i.e., model training) and the remaining 20% data (~40 thousand data points) were utilized in model validation (i.e., model testing) to optimize the model performance.

2.7 Model performance assessment

In the present study, the model performance was evaluated using two performance indices. These are: i) the coefficient of determination ($R^2$) and ii) the ratio of root-mean-square error (RMSE) over the standard deviation (SD) of observations (RSR) and are estimated as follows, Eq. (5a-5c):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_{i,m} - Y_{m,\text{mean}})(Y_{i,o} - Y_{o,\text{mean}})}{\sqrt{\sum_{i=1}^{N} (Y_{i,m} - Y_{m,\text{mean}})^2 \sum_{i=1}^{N} (Y_{i,o} - Y_{o,\text{mean}})^2}}$$  \hspace{1cm} (5a)

$$\text{RMSE} = \frac{1}{N} \sum_{i=1}^{N} (Y_{i,o} - Y_{i,m})^2$$  \hspace{1cm} (5b)

$$\text{SD} = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i,o} - Y_{o,\text{mean}})^2$$  \hspace{1cm} (5c)

Where $N$ is the number of data points, $Y_{i,o}$ is the observed response, $Y_{i,m}$ is the modeled/predicted response, $Y_{o,\text{mean}}$ and $Y_{m,\text{mean}}$ is the mean values of, respectively, $Y_{i,o}$ and $Y_{i,m}$.

The statistical measure $R^2$ is a widely used statistical score metric and often useful in model performance evaluation. It indicates the degree of correlation between predicted and observed values. In addition, it denotes the predictive power of a model. The value of $R^2$ ranges from 0 to 1, with a higher value representing the best fit model. Statistics provided a basis for the assessment of model calibration and validation and suggested a range of RSR values whether a model should be accepted or rejected [52]. The value of a ‘perfect to very good model’ ranges from 0 to 0.50, a ‘good model’ has a RSR value from 0.50 to 0.60, RSR value between 0.60 and 0.70 refers to a ‘satisfactory model’, and a model with RSR > 0.70 represents an ‘unacceptable model’.

3. Results

3.1. Correlations of DO with predictors

The linear correspondences of the response variable DO with different predictors were computed using the Pearson correlation coefficients $r$ (obtained from the transformed and standardized data) (Table 2). The correlation matrix showed that DO was strongly correlated ($r = -0.90$) with $T_w$ and was relatively weakly correlated ($r = -0.38$) with SC. These correlation results were statistically significant at 95% confidence intervals ($p$ value < 0.05). However, other predictors – $Q$, pH, TUR, and GH – showed very weak correlations ($r = -0.06$ to 0.16) with DO and were statistically not significant ($p$ value > 0.05). Further, the mutual correlations among the predictor variables indicated the presence of a moderate to strong multicollinearity in the data matrix. For example, $Q$
had a strong correlation \( (r = -0.98) \) with GH, moderate correlations \( (r = -0.61 \text{ and } -0.64, \text{ respectively}) \) with SC and TUR, and a weak correlation \( (r = -0.40) \) with pH. SC showed weak correlations \( (r = 0.42 \text{ and } 0.48, \text{ respectively}) \) with pH and \( T_w \) and a strong correlation \( (r = -0.73) \) with GH. Further, GH demonstrated a weak correlation \( (r = -0.45) \) with pH and a moderate correlation \( (r = 0.58) \) with TUR.

The presence of multicollinearity among predictors rationalized the application of VIMs to select important variables that have significant influences on the response variable.

### Table 2. Pearson correlation matrix among participatory variables.

|    | \( T_w \) | \( Q \) | SC | pH | TUR | GH | DO |
|----|----------|------|----|----|-----|----|----|
| \( T_w \) | 1 | -0.24 | 0.48 | 0.16 | -0.02 | -0.30 | -0.90 |
| \( Q \) | -0.24 | 1 | -0.61 | -0.40 | 0.64 | 0.98 | 0.11 |
| SC | 0.48 | -0.61 | 1 | 0.42 | -0.25 | -0.73 | -0.38 |
| pH | 0.16 | -0.40 | 0.42 | 1 | -0.25 | -0.45 | 0.07 |
| TUR | -0.02 | 0.64 | -0.25 | -0.25 | 1 | 0.58 | -0.06 |
| GH | -0.30 | 0.98 | -0.73 | -0.45 | 0.58 | 1 | 0.16 |
| DO | -0.90 | 0.11 | -0.38 | 0.07 | -0.06 | 0.16 | 1 |

Note. Moderate (~0.50 to <0.70) to strong (≥0.70) correlations were embolden and were statistically significant at 95\% confidence interval (p value < 0.05).

### 3.2. Important predictors based on VIMs

The variable importance score of each predictor was estimated and presented in a vertical bar chart (Figure 4). Water temperature was identified as the top-ranked (i.e., the most influential) predictor of DO, while turbidity was found at the bottom of the ranking (i.e., the least influential). In comparison with turbidity, water temperature and pH had approximately, respectively, 62- and 5-times stronger influence on DO dynamics. SC and Q approximately had 2 times stronger control over DO compared to turbidity. However, SC demonstrated a slightly stronger influence than that of Q (i.e., SC was 1.03 times stronger than Q). Further, both turbidity and gage height appeared to have a similar influence on model output. This led to the selection of 6 different combinations of predictors that were separately used in random forests model building as follows: i) \( T_w \) only; ii) \( T_w \) and pH only; iii) \( T_w \), pH, and SC only; iv) \( T_w \), pH, SC, and Q only; v) \( T_w \), pH, SC, Q, and GH only; and vi) all predictors. The selection of each scenario was based on the corresponding rank, following a descending order, of each variable.

![Figure 4. Individual predictor variable influence score on the response variable.](image-url)
3.3. Random forest model calibration and validation

Models for different combinations of predictors were developed using 80% of the data. Then each model was evaluated with the remaining 20% of the data. The calibration and validation results (Table 3 and Figure 5) for various combinations were described with a sequential manner.

Table 3. Random forest model statistics for different combinations of input variables.

| Predictors combination | Calibration $R^2$, RMSE | Validation $R^2$, RMSE |
|------------------------|--------------------------|------------------------|
| $T_w$                  | 0.851, 0.892             | 0.852, 0.892           |
| $T_w$ and pH           | 0.923, 0.643             | 0.912, 0.688           |
| $T_w$, pH, and SC      | 0.987, 0.259             | 0.951, 0.513           |
| $T_w$, pH, SC, and Q   | 0.998, 0.030             | 0.976, 0.357           |

![Graphs](image1.png)

**Figure 5.** The performance of the validated models. The model was developed considering: - a) $T_w$ only; b) $T_w$ and pH only; c) $T_w$, pH, and SC only; and d) $T_w$, pH, SC, and Q only. The red marked solid lines represent the regression lines. DO refers dissolved oxygen.

When the model was calibrated considering $T_w$ only, model statistics $R^2$ and RMSE were estimated as, respectively, 0.851 and 0.892 (Table 3). The corresponding validated $R^2$ and RMSE of the model output were estimated as, respectively, 0.852 and 0.892 (Table 3, Figure 5a). The calibrated
model $R^2$ and RMSE were estimated as, respectively, 0.923 and 0.643, when the model was constructed considering $T_w$ and pH. The corresponding validated $R^2$ and RMSE of the model output were estimated as, respectively, 0.912 and 0.688 (Figure 5b).

![Figure 6](preprints.org/doi:10.20944/preprints202004.0342.v1)

Figure 6. The comparison between measured dissolved oxygen (DO) vs. modeled (predicted) dissolved oxygen (DO) when the model was developed considering: a) $T_w$ only; b) $T_w$ and pH only; c) $T_w$, pH, and SC only; and d) $T_w$, pH, SC, and Q only.

The calibrated $R^2$ and RMSE of the response were estimated as, respectively, 0.987 and 0.259, considering $T_w$, pH, and SC as predictors in model development. Corresponding validated $R^2$ and RMSE of the predicted response were estimated as, respectively, 0.951 and 0.513 (Figure 5c).
calibrated $R^2$ and RMSE of the response were estimated as, respectively, 0.998 and 0.030, when $T_w$, pH, SC, and Q were selected as predictor variables in model formulation. Corresponding validated $R^2$ and RMSE of the predicted response were estimated as, respectively, 0.976 and 0.357 (Figure 5d). The RSR for different combinations (both in calibration and validation) varied within a range between 0.18 and 0.29, suggesting very good to perfect models.

It is worthy to note that the inclusion of four predictors ($T_w$, pH, SC, and Q) in calibrating the model was able to explain maximum 99.8% data variance. Therefore, other variables were not taken into account in model development. However, three predictors such as $T_w$, pH, and SC showed the optimal performance in model development by explaining 98.7% data variance. The comparison between DO used for validations and DO calculated from models also demonstrated that $T_w$, pH, and SC optimally performed in prediction (Figure 6).

3.4. Performance of the linear regression models

Performances of random forest models were further compared with the performances of MLR models. The model was developed using the training data and regression weights were estimated as follows, Eq. (6a-6d):

$$DO_{cal} = 13.90 - 0.23T_w \quad (6a)$$

$$DO_{cal} = 3.18 - 0.24T_w + 1.33pH \quad (6b)$$

$$DO_{cal} = 2.92 - 0.24T_w + 1.33pH - 0.002SC \quad (6c)$$

$$DO_{cal} = 3.91 - 0.24T_w + 1.33pH - 0.003SC + 6 \times 10^{-5}Q \quad (6d)$$

where, $DO_{cal} =$ DO calibrated value (mg/L); $T_w =$ water temperature ($^\circ$C); $pH =$ water pH; SC = specific conductance (μS/cm); $Q =$ river discharge (ft$^3$/s);

| Predictors combination | Calibration $R^2$, RMSE | Validation $R^2$, RMSE |
|------------------------|-------------------------|------------------------|
| $T_w$                  | 0.764, 1.011            | 0.812, 1.003           |
| $T_w$ and pH           | 0.833, 0.874            | 0.860, 0.865           |
| $T_w$, pH, and SC      | 0.834, 0.872            | 0.861, 0.865           |
| $T_w$, pH, SC, and Q   | 0.834, 0.874            | 0.863, 0.856           |

The calibrated $R^2$ and RMSE for Eq. (6a) were estimated as, respectively, 0.764 and 1.011 for Eq. (6a), 0.833 and 0.874 for Eq. (6b), 0.834 and 0.872 for Eq. (6c), and 0.834 and 0.874 for Eq. (6d) (Table 4). The RSR values ranged between 0.39 and 0.41 across these three calibrated models. The regression coefficients were used to test the remaining 20% of the data.

The validated $R^2$ and RMSE were estimated as, respectively, 0.812 and 1.003 for Eq. (6a), 0.860 and 0.865 for Eq. (6b), 0.861 and 0.865 for Eq. (6c), and 0.863 and 0.856 for Eq. (6d). The RSR values varied between 0.32 and 0.36 across these four validated models. For calibrated models, Eq. (6b-6d), the condition number were estimated as a three-order of magnitude (i.e., $10^3$) – a proof of strong multicollinearity in the data matrix.

In comparison, the random forest method performed better by explaining more data variance ($R^2 = 0.851-0.998$) and demonstrated very good model accuracy (RSR = 0.18-0.39) both in model calibration as well as model validation. Like random forests, the model development for different calibration scenarios was also limited to only four predictor variables such as $T_w$, pH, SC, and Q as
an addition of other predictor variables did not significantly improve the model performances. The satisfactory performance was achieved by considering only three predictors (T, pH, and SC).

4. Discussion

The present study effectively identified the major drivers of DO by estimating the relative influence of each attribute employing the Gini decrease index [48]. This information would be potentially useful for water quality managers in future research and crucial decision making, especially in priority setting in pollution control of streams and rivers.

The comprehensive analyses across all input scenarios suggest that the random forest method emerged as a powerful data-driven tool to estimate DO concentrations, manifested by the model statistics, over the conventional MLR method. All combinations of random forests showed an improved performance (both in model formulation and testing) compared to the performance of MLR method. Although MLR models also showed good calibration and validation accuracy and explained up to ~86% of the total data variance, possible cross-correlations (Table 2) among participatory variables left the MLR technique an unattractive one due to unstable estimates [53] and high condition numbers (i.e., ill-conditioning of linear models).

The study results show a consistency with other studies. For example, Heddam [54] in a separate study demonstrated that water temperature and pH can predict DO levels with an enough level of confidence. The results and insight from this study can also offer practical benefits. We would be able to replicate established models in similar regions to predict real-time DO levels with less input data. Although four predictors (T, pH, SC, and Q) explained the maximum data variance, it is recommended to use three predictors (T, pH, and SC) to estimate DO levels which are relatively easy to measure in field. Therefore, it further indicates the parsimony of input variables. The inclusion of Q in model building only increased 1.1% of the data variance. However, it is often difficult to have a continuous measurement of discharge due to the high uncertainty (i.e., sudden flood) associated with it, especially for larger rivers; therefore, can be avoided. Further, only T and pH can be used to efficiently predict DO level (calibrated model explained ~91% data variance), particularly in small streams where DO is not properly monitored. Even with T only, the model (explained ~85% data variance) can be applied to monitor DO levels with an acceptable level of confidence. At unmonitored sites, the water temperature can be estimated from the available air temperature [55]. This would be particularly useful when a river/stream is inaccessible due to complex geography or high labor expenditure for taking measurements.

A well-designed river water quality monitoring program is a requisite to keep track of DO levels in streams in order to protect valuable aquatic life and their essential habitat [56]. While water monitoring to serve myriad intents is generally well defined, the evaluation of water quality is often problematic due to many input parameters [57,58]. The considerable cost reduction due to the parsimony of required field data would help ameliorate water resources management by taking prompt decisions.

Despite the potential implication, the developed models require further careful testing in various locations with newer set of data. The models potentially hold an empirical basis and are based on the data between 2007 and 2019 obtained from the USGS data repository system. It is likely that the measured values under certain circumstances were not exactly reported due to measurement errors and/ personnel unawareness. Data also contain missing values at times, especially between 2015 – 2016, and were filtered following an outlier removal technique that might have inserted errors in results. Moreover, the study area was dominated by a vegetative land cover (i.e., mainly deciduous forest). The predominance of other land use activities (i.e., agriculture dominated/developed area) could bring different findings. Subjected to the availability of more predictor variables, across a diverse environment and land use gradient, with no missing data, the present model should further be calibrated and tested accordingly to observe the overall local or global robustness. Nevertheless, this simple data-driven model would be of enormous help to develop strategic framework in managing ecosystems across the U.S. that require immediate attention, as mandated by the U.S. Clean Water Act [59], and different parts of the world.
5. Conclusions

A random forest machine learning technique has been successfully applied to predict DO concentrations of the South Branch Potomac River near Springfield, WV. The variable important matrix resulted in 6 distinct combinations of predictors that were separately employed in the development of random forests as follows: i) $T_w$ only; ii) $T_w$ and pH only; iii) $T_w$, pH, and SC only; iv) $T_w$, pH, SC, and Q only; v) $T_w$, pH, SC, Q, and GH only; and vi) all predictors. The model was trained and validated using, respectively, 80% and 20% of the data. The model statistics ($\text{R}^2$ and RSR) were estimated in each case and the first four input combinations were adequate for model building.

Results showed that $T_w$ and pH can efficiently predict DO concentration, with 91-92% data variance, across the calibration and validation phases. The best model performance (i.e., explained maximum 99.8% data variance) in calibration stage can be achieved by considering four predictors such as $T_w$, pH, SC, and Q as input, decreasing in rank. However, the recommended input variables, to model formulation, are $T_w$, pH, and SC given their satisfactory performance (i.e., explained 98.7% data variance). Estimated RSR further suggested very good to perfect (i.e., high efficiency) models. On the contrary, traditional MLR models performed with less accuracy, both in model building and testing, by explaining only 76-86% data variance across different input scenarios. Further, potential cross-correlations among predictor variables and high condition numbers indicated the potential bias in traditional MLR model estimations. Therefore, the random forest model evolved as a powerful data-driven tool to estimate DO concentrations which requires less input data but demonstrates an improved performance with a higher level of accuracy.

Although developed models, using a random forest, presented better performances in predicting DO levels, when compared with a MLR method, there is still a room for the further assessment. Models need to be further calibrated and validated with a new data matrix with no missing data encompassing more predictor variables across diverse climates and management gradients (i.e., various land-cover dominated catchments). However, the developed models can still direct in future research. We would be able to predict DO concentrations with a minimal number of input variables. This would be particularly beneficial for streams/rivers where the water quality monitoring is not properly maintained due to geographical access constraints or high labor cost. The potential of parsimonious model development, with a significant cost reduction in data collection and processing, is therefore expected to guide water resource managers to take strategic and prompt measures towards achieving a healthy ecosystem across the continental U.S., as mandated by the Clean Water Act, and beyond.

Author Contributions: Mohammad Hafez Ahmed is the whole contributor of this manuscript.

Funding: This research received no external funding.

Acknowledgments: I would like to thank all USGS scientists and administrations for allowing to use the water quality data and make it openly accessible. I also thank two anonymous reviewers for their constructive comments that helped significantly improve the manuscript.

Conflicts of Interest: The author declares no conflict of interest.
### Appendix A

**Figure A1.** Time-series data of water quality variables from the year 2007 to 2019. $T_w$, $Q$, SC, pH, TUR, GH, represents, respectively, water temperature (°C), discharge ($m^3/s \times 0.028$), specific conductance (uS/m $\times 0.01$), water pH, turbidity (FNU), and gage height ($m \times 0.308$).

### References

1. Lintern, A.; Webb, J.A.; Ryu, D.; Liu, S.; Bende-Michl, U.; Waters, D.; Leahy, P.; Wilson, P.; Western, A.W. Key factors influencing differences in stream water quality across space. *WIREs Water* **2018**, *5*, 1260, doi:10.1002/wat2.1260.

2. Kroeze, C.; Gabbert, S.; Hofstra, N.; Koelmans, A.A.; Li, A.; Löhr, A.; Ludwig, F.; Strokal, M.; Verburg, C.; Vermeulen, L.; et al. Global modelling of surface water quality: a multi-pollutant approach. *Curr. Opin. Environ. Sustain.* **2016**, *23*, 35–45, doi:10.1016/J.COSUST.2016.11.014.

3. Gao, C.; Zhang, T. Eutrophication in a Chinese context: understanding various physical and socio-economic aspects. *Ambio* **2010**, *39*, 385–93, doi:10.1007/s13280-010-0040-5.

4. Chapra, S. *Surface water-quality modeling*; Newyork: McGraw-Hill Companies, Inc., 2008;

5. Wen, X.; Fang, J.; Diao, M.; Zhang, C. Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. *Environ. Monit. Assess.* **2013**, *185*, 4361–4371, doi:10.1007/s10661-012-2874-8.

6. Chin, D.A. *Water-Quality Engineering in Natural Systems*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006; ISBN 9780471784555.

7. Vellidis, G.; Barnes, P.; Bosch, D.D.; Cathey, A.M. Mathematical Simulation Tools for Developing Dissolved Oxygen TMDLs. **2006**.
8. Loucks, D.P.; van Beek, E. Water Quality Modeling and Prediction. In Water Resource Systems Planning and Management; Springer International Publishing: Cham, 2017; pp. 417–467.

9. Sánchez, E.; Colmenarejo, M.F.; Vicente, J.; Rubio, A.; García, M.G.; Travieso, L.; Borja, R. Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. Ecol. Indic. 2007, 7, 315–328, doi:10.1016/j.ecolind.2006.02.005.

10. Rajwa, K.; Agnieszka, B.; Rowiński, R.J.; M., P. Dissolved oxygen and water temperature dynamics in lowland rivers over various timescales. J. Hydrol. Hydromech. 2015, 63, 353–363, doi:10.1515/johh-2015-0041.

11. Ranković, V.; Radulović, J.; Radojević, I.; Ostojić, A.; Čomić, L. Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. Ecol. Model. 2010, 221, 1239–1244, doi:10.1016/j.ecolmodel.2009.12.023.

12. Streeter, H.W.; Phelps, E.B. A Study of the Pollution and Natural Purification of the Ohio Rivers.; U.S. Public Health Service Bulletin, 1925;

13. Cox, B.A. A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. Sci. Total Environ. 2003, 314–316, 335–377, doi:10.1016/S0048-9697(03)00063-9.

14. Kannel, P.R.; Kanel, S.R.; Lee, S.; Lee, Y.S.; Gan, T.Y. A Review of Public Domain Water Quality Models for Simulating Dissolved Oxygen in Rivers and Streams. Environ. Model. Assess. 2011, 16, 183–204.

15. Li, G. Stream temperature and dissolved oxygen modeling in the Lower Flint River Basin. PhD Dissertation, University of Georgia, Athens, GA, 2006.

16. Antonopoulos, V.Z.; Gianniou, S.K. Simulation of water temperature and dissolved oxygen distribution in Lake Vegoritis, Greece. Ecol. Model. 2003, 160, 39–53, doi:10.1016/S0304-3800(02)00286-7.

17. Kuo, J.T.; Lung, W.S.; Yang, C.P.; Liu, W.C.; Yang, M. Der; Tang, T.S. Eutrophication modelling of reservoirs in Taiwan. Environ. Model. Softw. 2006, 21, 829–844, doi:10.1016/j.envsoft.2005.03.006.

18. Ay, M.; Kisi, O. Modeling of Dissolved Oxygen Concentration Using Different Neural Network Techniques in Foundation Creek, El Paso County, Colorado. J. Environ. Eng. 2012, 138, 654–662, doi:10.1061/(ASCE)EE.1943-7870.0000511.

19. Heddam, S. Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. Environ. Technol. U. K. 2014, 35, 1650–1657, doi:10.1080/09593330.2013.878396.

20. Zhang, Y.-F.; Fitch, P.; Thorburn, P.J. Predicting the Trend of Dissolved Oxygen Based on the kPCA-RNN Model. Water 2020, 12, 585, doi:10.3390/w12020585.

21. Altunkaynak, A.; Özger, M.; Çakmakçı, M. Fuzzy logic modeling of the dissolved oxygen fluctuations in Golden Horn. Ecol. Model. 2005, 189, 436–446, doi:10.1016/J.ECOLMODEL.2005.03.007.

22. Heddam, S. Modeling hourly dissolved oxygen concentration (DO) using two different adaptive neuro-fuzzy inference systems (ANFIS): a comparative study. Environ. Monit. Assess. 2014, 186, 597–619, doi:10.1007/s10661-013-3402-1.
23. Ay, M.; Kişi, Ö. Estimation of dissolved oxygen by using neural networks and neuro fuzzy computing techniques. *KSCE J. Civ. Eng.* 2017, 21, 1631–1639, doi:10.1007/s12205-016-0728-6.

24. Chen, W.-B.; Liu, W.-C. Artificial neural network modeling of dissolved oxygen in reservoir. *Environ. Monit. Assess.* 2014, 186, 1203–1217, doi:10.1007/s10661-013-3450-3.

25. Nemati, S.; Fazelifard, M.H.; Terzi, Ö.; Ghorbani, M.A. Estimation of dissolved oxygen using data-driven techniques in the Tai Po River, Hong Kong. *Environ. Earth Sci.* 2015, 74, 4065–4073, doi:10.1007/s12665-015-4450-3.

26. Tarmizi, A.; Ahmed, A.N.; El-Shafie, A. Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River. *Middle-East J. Sci. Res.* 2014, 21, 2182–2188, doi:10.5829/idosi.mejsr.2014.21.11.21844.

27. Ji, X.; Shang, X.; Dahlgren, R.A.; Zhang, M. Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China. *Environ. Sci. Pollut. Res.* 2017, 24, 16062–16076, doi:10.1007/s11356-017-9243-7.

28. Heddam, S.; Kisi, O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol.* 2018, 559, 499–509, doi:10.1016/j.jhydrol.2018.02.061.

29. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood hazard risk assessment model based on random forest. *J. Hydrol.* 2015, 527, 1130–1141, doi:10.1016/j.jhydrol.2015.06.008.

30. Granata, F.; Saroli, M.; de Marinis, G.; Gargano, R. Machine Learning Models for Spring Discharge Forecasting Available online: https://www.hindawi.com/journals/geofluids/2018/8328167/ (accessed on Apr 15, 2020).

31. Granata, F. Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agric. Water Manag.* 2019, 217, 303–315, doi:10.1016/j.agwat.2019.03.015.

32. Granata, F.; Gargano, R.; de Marinis, G. Artificial intelligence based approaches to evaluate actual evapotranspiration in wetlands. *Sci. Total Environ.* 2020, 703, 135653, doi:10.1016/j.scitotenv.2019.135653.

33. Naghibi, S.A.; Pourghasemi, H.R.; Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess.* 2015, 188, 44, doi:10.1007/s10661-015-5049-6.

34. Chen, X.-W.; Liu, M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics* 2005, 21, 4394–4400, doi:10.1093/bioinformatics/bti721.

35. Chagas, C. da S.; de Carvalho Junior, W.; Bhering, S.B.; Calderano Filho, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *CATENA* 2016, 139, 232–240, doi:10.1016/J.CATENA.2016.01.001.
36. Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674–2674, doi:10.3390/s18082674.

37. Wang, Z.; Wang, Y.; Zeng, R.; Srinivasan, R.S.; Ahrentzen, S. Random Forest based hourly building energy prediction. *Energy Build.* **2018**, *171*, 11–25, doi:10.1016/J.ENERBUILD.2018.04.008.

38. The Water Quality Portal Available online: https://www.waterqualitydata.us/ (accessed on Sep 13, 2019).

39. Multi-Resolution Land Characteristics (MRLC) Consortium | Multi-Resolution Land Characteristics (MRLC) Consortium Available online: https://www.mrlc.gov/ (accessed on Mar 20, 2020).

40. Tromans, D. Temperature and pressure dependent solubility of oxygen in water: a thermodynamic analysis. *Hydrometallurgy* **1998**, *48*, 327–342, doi:10.1016/S0304-386X(98)00007-3.

41. Schaefer, S.C.; Hollibaugh, J.T. Temperature Decouples Ammonium and Nitrite Oxidation in Coastal Waters. *Environ. Sci. Technol.* **2017**, *51*, 3157–3164, doi:10.1021/acs.est.6b03483.

42. Chow, V. Te; Maidment, D.R.; Mays, L.W. *Applied hydrology*; McGraw-Hill, 1988; ISBN 0070108102.

43. Massey, F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78, doi:10.1080/01621459.1951.10500769.

44. Box, G.E.P.; Cox, D.R. An Analysis of Transformations. Journal of the Royal Statistical Society.Series B (Methodological): 211-252 1964, 211–252.

45. Carrascal, L.M.; Galván, I.; Gordo, O. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* **2009**, *118*, 681–690, doi:10.1111/j.1600-0706.2008.16881.x.

46. Mannina, G.; Viviani, G. Water quality modelling for ephemeral rivers: Model development and parameter assessment. *J. Hydrol.* **2010**, *393*, 186–196, doi:10.1016/j.jhydrol.2010.08.015.

47. Muttil, N.; Chau, K.W. Machine-learning paradigms for selecting ecologically significant input variables. *Eng. Appl. Artif. Intell.* **2007**, *20*, 735–744, doi:10.1016/j.engappai.2006.11.016.

48. Ai, F.; Bin, J.; Zhang, Z.; Huang, J.; Wang, J.; Liang, Y.; Yu, L.; Yang, Z. Application of random forests to select premium quality vegetable oils by their fatty acid composition. *Food Chem.* **2014**, *143*, 472–478, doi:10.1016/j.foodchem.2013.08.013.

49. Boulesteix, A.-L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507, doi:10.1002/widm.1072.

50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.

51. Kuhn, M.; Johnson, K. *Applied Predictive Modeling [Hardcover]*; Newyork, 2013; ISBN 978-1-4614-6848-6.
52. Moriasi, D.N.; Arnold, J.G.; Liew, M.W. Van; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Trans. ASABE* **2007**, *50*, 885–900, doi:10.13031/2013.23153.

53. Öztürk, F.; Akdeniz, F. *Ill-conditioning and multicollinearity*; 2000; Vol. 321:.

54. Heddam, S. Simultaneous modelling and forecasting of hourly dissolved oxygen concentration (DO) using radial basis function neural network (RBFNN) based approach: a case study from the Klamath River, Oregon, USA. *Model. Earth Syst. Environ.* **2016**, *2*, doi:10.1007/s40808-016-0197-4.

55. Mohseni, O.; Stefan, H.G. Stream temperature/air temperature relationship: A physical interpretation. *J. Hydrol.* **1999**, *218*, 128–141, doi:10.1016/S0022-1694(99)00034-7.

56. Kannel, P.R.; Lee, S.; Lee, Y.S.; Kanel, S.R.; Khan, S.P. Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. *Environ. Monit. Assess.* **2007**, *132*, 93–110, doi:10.1007/s10661-006-9505-1.

57. Chapman, D. *Water Quality Assessments-A Guide to Use of Biota, Sediments and Water in Environmental Monitoring-Second Edition*; 1992; pp. 585–585:.

58. Dixon, W.; Chiswell, B. Review of aquatic monitoring program design. *Water Res.* **1996**, *30*, 1935–1948, doi:10.1016/0043-1354(96)00087-5.

59. Federal Water Pollution Control Act *Federal Water Pollution Control Act, as amended through P.L.107–303, November 27, 2002, in U.S. Congress, Washington, D. C.*; 2002;