Prediction of effluent arsenic concentration of wastewater treatment plants using machine learning and kriging-based models

Mohammad Zounemat-Kermani1,2 · Meysam Alizamir3 · Behrooz Keshtega4 · Okke Batelaan5 · Reinhard Hinkelmann6

Received: 23 March 2021 / Accepted: 3 October 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
This study evaluates the potential of kriging-based (kriging and kriging-logistic) and machine learning models (MARS, GBRT, and ANN) in predicting the effluent arsenic concentration of a wastewater treatment plant. Two distinct input combination scenarios were established, using seven quantitative and qualitative independent influent variables. In the first scenario, all of the seven independent variables were taken into account for constructing the data-driven models. For the second input scenario, the forward selection k-fold cross-validation method was employed to select effective explanatory influent parameters. The results obtained from both input scenarios show that the kriging-logistic and machine learning models are effective and robust. However, using the feature selection procedure in the second scenario not only made the architecture of the model simpler and more effective, but also enhanced the performance of the developed models (e.g., around 7.8% performance enhancement of the RMSE). Although the standard kriging method provided the least good predictive results (RMSE = 0.18 ug/l and NSE=0.75), it was revealed that the kriging-logistic method gave the best performance among the applied models (RMSE = 0.11 ug/l and NSE=0.90).

Keywords Kriging-logistic · Artificial neural networks · Gradient boosted regression trees · Multivariate adaptive regression splines · Arsenic mitigation

Nomenclature
ANFIS Adaptive neuro-fuzzy inference system
ANN Artificial neural network
As Arsenic
BF Basis function
BOD Biochemical oxygen demand
COD Chemical oxygen demand
GBRT Gradient boosted regression trees
GCV Generalized cross-validation
GM Grey model
GRI Geometric reliability index
MAE Mean absolute error
MAGE Mean absolute geometric error
MARS Multivariate adaptive regression splines
ML Machine learning
NSE Nash-Sutcliffe efficiency
PCC Pearson’s correlation coefficient
RMSE Root mean square error

Highlights • Concentration of dissolved arsenic in wastewater is predicted.
• Prediction efficiency of two different data-driven strategies is evaluated.
• Machine learning (MARS, GBRT, and ANN), kriging, and kriging-logistic methods are used.
• The kriging-logistic method gave the best performance.

Responsibility Editor: Marcus Schulz

Mohammad Zounemat-Kermani zounemat@uk.ac.ir

1 Department of Water Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
2 Center of Excellence in Hydroinformatics, University of Tabriz, 29 Bahman Ave, Tabriz, Iran
3 Department of Civil Engineering, Hamedan branch, Islamic Azad University, Hamedan, Iran
4 Department of Civil Engineering, Faculty of Engineering, University of Zabol, P.B. 9861335-856, Zabol, Iran
5 National Centre for Groundwater Research and Training, College of Science and Engineering, Flinders University, Adelaide, Australia
6 Chair of Water Resources Management and Modeling of Hydrosystems, Technische Universität Berlin, Berlin, Germany

Published online: 05 November 2021
**Introduction**

Most municipal and hospital wastewater treatment plants (WWTPs) have regular effluent inspections in order to meet regulation standards. However, due to high laboratory analysis costs, human mistakes, and unprecedented technical problems, it is not always possible to rely on regular measurements. Under these conditions, the effluent quality cannot always be appropriately predicted using statistical and numerical methods. In addition, the water quality of a WWTP is sensitive to parameters such as water temperature, pH, influent discharge, and contaminants. This is due to the fact that in WWTPs, the wastewater is biologically treated by metabolism processes resulting from the activity of microorganisms. Hence, the treatment process follows a complex and highly nonlinear trend influenced by several known and unknown independent parameters (Guo et al. 2015). In that sense, having a reliable trained mathematical system, either for data prediction in case of having missing data or checking those measured data with ambiguity in the measurement process, would be of great value.

Models based on data mining approaches, such as machine learning (ML) and data mining models, have already proven to be appropriate and effective tools in dealing with various environmental engineering problems (Haghiabi et al. 2018; Zounemat-Kermani et al. 2020; Şimsek and Alkay 2020; Yetis et al. 2021). Given their architecture, ML models are capable of predicting nonlinear systems’ output value(s) just relying on the input parameters after the ML models have been trained and evaluated.

Due to the complexity of the wastewater treatment processes, several researchers have already attempted to evaluate the effectiveness of machine learning models in modeling and predicting water quality in WWTP effluent (Qin et al. 2012; Zounemat-Kermani et al. 2019; Bernardelli et al. 2020; Icke et al. 2020). Pai et al. (2007) used grey model (GM) and artificial neural network (ANN) to predict chemical oxygen demand (COD) and suspended solids (SS) in a hospital effluent from a wastewater treatment plant. They confirmed the suitability of ANN application in the prediction process. In another similar study, Pai et al. (2009) developed machine learning methods including adaptive neuro-fuzzy inference system (ANFIS) and ANN for the effluent prediction of SS and COD of a hospital wastewater treatment plant. It was found that the employed ANFIS model acted better in the prediction of the effluent parameters. Guo et al. (2015) employed ANNs and support vector machines (SVMs) to predict the effluent concentration in a WWTP in Korea. Both models performed acceptably in the prediction procedure; however, it was found that the SVM model surpassed the ANN in terms of models’ accuracy. Sharafati et al. (2020) applied ensemble ML models, including the AdaBoost regression, gradient boost regression, and random forest regression, in predicting the WWTP effluent water quality parameters, such as total dissolved solids (TDS), COD, and biochemical oxygen demand (BOD). The AdaBoost and gradient boost regression methods showed a better performance in the prediction process of TDS and COD, respectively. Granata et al. (2017) attempted to model and estimate the wastewater quality such as BOD, COD, TSS, and TDS indicators using support vector regression (SVR) and regression tree (RT) as black-box indirect modeling. They reported successful application of both machine learning models signifying that the SVR model acted better than the RT model. Ansari et al. (2018) modeled and predicted the influent flow rate of a sewage treatment plant using a stochastic, nonlinear autoregressive network and SVM model. The results of the study demonstrated that the SVM model was successful in modeling peak inflow. Fard et al. (2020) trained a multilayer perceptron ANN by the Levenberg-Marquardt algorithm and employed a genetic algorithm to predict the total COD, ammonia, phosphate, and turbidity of the effluent of a slaughterhouse WWTP. The findings indicated that the ANN trained by the genetic algorithm performed better than the standard ANN.

Arsenic is a toxic heavy metal and is seen as being synonymous with severe contamination of water bodies. The World Health Organization recommends that arsenic concentrations in drinking water should be below 10 ppb (parts per billion) (Yapıcıoğlu et al. 2020). Arsenic can be found in aquatic systems in its inorganic form resulting from the dissolution of solid phases like arsenic anhydride (As2O3), arsenolite (As2O5), and realgar (AsS2) (Harper and Kingham 1992). Nowadays, high arsenic concentrations in surface waters and groundwaters are a worldwide problem; hence, this study focuses on the modeling of this important parameter in the effluent of WWTP’s.

The reason for the presence of arsenic contamination in wastewater has been discussed in some research papers (Harper and Kingham 1992; Komorowicz and Baralkiewicz 2016). According to the literature review, arsenic may be found in wastewaters for the following two main reasons.

1) Contamination of the main water resources (e.g., the existence of arsenic in groundwater used as a source of drinking water)
Human activities that make the surface/drinking water polluted (e.g., ore mining, fossil fuel combustion, preparation of arsenic-containing drugs, and the use of pesticides)

Different strategies have been employed to remove arsenic or diminish concentration levels in drinking water, industrial, and wastewater effluent (Sheikhi et al. 2021). The most used technologies for removing or mitigating heavy metals from contaminated water are membrane distillation (Yapıcıoğlu 2020), electrocoagulation (Yapıcıoğlu 2018a), and biosorption (Agarwal et al. 2020). In this regard, arsenic removal technologies in WWTPs and other facilities cover a wide range of methods, such as oxidation (Zhang et al. 2018; Yapıcıoğlu 2018b), coagulation and softening (Ge et al. 2020), sorption and ion-exchange techniques (Zhou et al. 2017), synergistic reduction (Wang et al. 2020), membrane technologies (Hubadillah et al. 2019), and bioremediation (Plewniak et al. 2018). Although there are no related works in using ML models to model/predict effluent arsenic in WWTPs, several chemical and environmental engineering researchers developed ML approaches to model the removal process of arsenic. Instances include using ANNs in the prediction of the arsenic removal from aqueous solution by living cells of Bacillus cereus biomass (Giri et al. 2011), using an ANFIS model to simulate the removal of arsenic and chromium from water (Mandal et al. 2015), and using ANNs to model the removal of aquatic arsenic species by indigenous microbes (Altowayti et al. 2020).

This study aims to predict the effluent arsenic concentration in WWTPs based on several influent water quality parameters, such as influent arsenic values, flow, BOD, TSS, pH, temperature, and turbidity, using data-driven techniques, including kriging, logistic-kriging, multilayer perceptron ANN, gradient boosted regression trees (GBRT), and multivariate adaptive regression splines (MARS). To the best of the authors’ knowledge, previous studies have not implemented statistical and ML methods to model effluent arsenic concentration in WWTPs. Moreover, pertinent studies in the field of arsenic removal modeling using ML methods are confined to conventional and simple soft computing methods such as ANNs and ANFIS. Accordingly, the present study’s contribution not only comprises the specific case study problem but also challenges the capability of five different data mining statistical (kriging and kriging-logistic) and ML (ANN, GBRT, and MARS) methods in the prediction procedure. It should be noted that in this work, the flexibility of non-linearity for input data of the developed kriging model is enhanced using the logistic map. This logistic map has a nonlinear scalar factor, which is determined for each input variable to control their nonlinearity.

Case study and dataset description

Ten years of monthly quantity and quality influent (input parameters) and effluent (target parameter) have been used for the development of the models. The input vector consists of influent variables, including arsenic concentration, flow, BOD, total sediment solids (TSS), pH, water temperature, and turbidity. On the other hand, the output (target) parameter is the effluent arsenic concentration. All datasets were extracted from the annual pretreatment reports for the Point Loma Wastewater Treatment Plant, San Diego City, from 2009 to 2019. The Point Loma WWTP opened in 1963 and treats almost 660,000 m³ of wastewater per day. The reports include a summary of pretreatment program activities of the metropolitan sewerage system, including the Point Loma WWTP, the South Bay Water Reclamation Plant, and the North City Water Reclamation Plant (Pretreatment Annual Report for the Point Loma POTW, 2009–2019). The general process for the wastewater treatment and pollutant removal at Point Loma plant consists of several stages, including (i) influent screens for removing large debris, (ii) grit chamber for settling heavy debris, such as sand; (iii) sedimentation tanks for removing heavy particles, organic solids, and scum; (iv) anaerobic digesters for converting organic matter to methane gas, carbon dioxide, and biosolids; and (iv) waste gas burners and cogeneration gas utilization facility.

Figure 1 illustrates the location of the Point Loma WWTP. Table 1 summarizes the statistical characteristics of the data used in this study. One interesting fact about the mean values presented in Table 1 is that, in general, the WWTP has mitigated the effluent arsenic by 30% (from 1.42 to 0.99 ug/L).

Besides, based on the highest Pearson’s correlation coefficient provided in Table 1, Figure 2 depicts the annual variations of the influent and effluent arsenic concentration using boxplots. As shown in Fig. 2, the effluent arsenic follows the general variational trend of the influent arsenic values. Moreover, it can also be realized that both influent and effluent arsenic concentrations have recorded new maximum values in 2018 and 2019.

Methods

This section describes the methodology and philosophy behind developing the GBRT, MARS, and kriging-based models. The results of these models are later compared with the conventional and well-known perceptron ANN, as a benchmark model, for evaluating the capability of the applied ML and kriging-based models. It should be noted
that for details on the perceptron ANN method, we refer to numerous available papers and reports explaining the perceptron ANN, such as Gardner and Dorling (1998) and Zounemat-Kermani et al. (2009).

**Gradient boosted regression trees (GBRT)**

GBRT is one of the most widely applied techniques for regression-based problems (Friedman 2001; Zounemat-Kermani et al. 2021). The core concept of the GBRT is based on improving the regression tree (RT) performance by employing the most explanatory input variables to reach the most accurate corresponding target. Generally, the RTs generate several leaf nodes, by mapping the predictor features to target values, which make several advantages like easy to find the interaction between input and response parameters, comprehensive structure to analyze the obtained results, and finally, computationally efficient (Song et al. 2019; Samadi et al. 2021). In the GBRT approach, generating initial RTs is carried out as the first step. Then, the RTs are trained through multiple iterations to minimize residual values. The forward stage format of the GBRT can be described as follows:

\[ F_m(x) = F_{m-1}(x) + h_m(x) \]  

\( \text{(1)} \)
where $F(x)$ and $h(x)$ denote the GBRT model and basis function, respectively. Also, $m$ and $x$ represent boosting iteration and input parameters. By finding the optimum value for $h(x)$, response values can be obtained using the following equation:

$$F_m(x_{i,t}) = F_{m-1}(x_{i,t}) + h_m(x_{i,t}) = Y_{i,t+k}$$

(2)

where $Y_{i,t+k}$ indicates the target value.

After each iteration, a new RT is fitted to reach a minimum error for the training phase. To improve the generalization performance by avoiding the over-fitting problem, the GBRT model applies the learning rate factor ($\eta$) to set the contribution of each RT in the final model by considering the number of iterations (Hastie et al. 2009; Song et al. 2019; Alizamir et al. 2020a). After employing this factor, the main equation can be presented as follows (Kim et al. 2020):

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad \eta[0;1]$$

(3)

Smaller values of $\eta$ helps the model to perform more accurate during the testing phase; however, in this case, the model also needs more iterations. In this study, parameters of the GBRT model were tuned based on a trial-and-error process to minimize differences between the target and actual values. In addition, the shrinkage factor, subsampling factor, and maximum tree depth were set to 0.015, 0.25, and 5, respectively. Also, 150 iterations and the squared loss functions were applied for the training process. More detail and parameters for the prediction of complex problems such as hydrological fields (Heddam et al. 2020), structural design optimization (Sakata et al. 2003; Huang et al. 2006), energy (Keshtegar et al. 2018; Zhang et al. 2020), and structural reliability analysis (Yun et al. 2018; Jiang et al., 2018b; Zhang et al. 2020; Jiang et al. 2020; Lu et al. 2019; Xiao et al. 2020) due to its ability in handling the high-nonlinear relations.

**Multivariate adaptive regression splines (MARS)**

MARS was proposed by Friedman (1991) as a robust non-parametric statistical tool to extract the relationship between the response and input parameters. The MARS model constructs a linear basis function by dividing the training dataset into splines and slope values. After fitting a regression equation to the training dataset, basis functions (BFs) and knots can be obtained (Heddam and Kisi 2018). The MARS algorithm achieves more error reduction by employing a forward phase for improving the location of each knot and its corresponding pair of the BFs. Also, the backward phase improves the performance of the MARS models by deleting the redundant BFs that have negligible effects on the model using the generalized cross-validation (GCV). The general expression of a MARS model can be presented as follows (Zhang and Goh 2016; Alizamir et al. 2020b):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(X)$$

(4)

$$\text{GCV}(M) = \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{(1 - \frac{G(M)}{n})^2}$$

(5)

where $f(x)$ stands for the dependent variable. $\lambda_m$ and $\beta$ are respectively a spline function and a constant coefficient calculated by using the least square methodology. Also, a linear function in the MARS model can be defined at value $t$ as follows:

$$\max(0, x - t) = \begin{cases} x - t, & \text{if } x \geq t \\ 0, & \text{otherwise} \end{cases}$$

(6)

Finally, the GCV value indicated the best-fit model. For this study, the GCV value and highest degree of interactions in the final model were equal to 0.0128 and 1, respectively. More details regarding MARS are available in Friedman (1991) and Friedman and Roosen (1995).

**Kriging-based methods**

The kriging-based surrogate model was introduced by Krig and Matheron (1952), while Matheron (1963) mathematically formulated the kriging model. Commonly, the kriging models can be applied for the prediction of complex problems such as hydrological fields (Heddam et al. 2020), structural design optimization (Sakata et al. 2003; Huang et al. 2006), energy (Keshtegar et al. 2018; Zhang et al. 2020), and structural reliability analysis (Yun et al. 2018; Jiang et al., 2018b; Zhang et al. 2020; Jiang et al. 2020; Lu et al. 2019; Xiao et al. 2020) due to its ability in handling the high-nonlinear relations.

In the standard and conventional kriging model, two terms are used for the interpolation of a mathematical relation as follows:

$$P(X) = G(X)^T \beta + r(X)^T \gamma$$

(7)

In the first term of Equation (7), $G(X)$ and $\beta$ respectively denote the basis function (BF) and unknown coefficient vectors. The nonlinearity of the predicted model is controlled by the nonlinear form of function $G(X)$. For accurate predictions with highest tendency and agreement, the BF mathematical form should be selected on the tended line of the prediction model. The second term represents the stochastic term, where:

$$\gamma = R^{-1}(O - G^T \beta)$$

(8)

and

$$r(X) = [R(\theta, X_1, X), R(\theta, X_2, X), \ldots, R(\theta, X_n, X)]^T$$

(9)

In Equations (8) and (9), $O$ represents the observed data in the training phase. $R$ denotes the correlation function,
which is given a stationary Gaussian function with the input of \(X_p, X_q\) (\(p, q = 1, 2, ..., m\)) (where \(p\)-th and \(q\)-th are input samples for \(m\) data points) and correlation parameter \(\theta\). The parameter \(\theta\) is determined by using the maximum likelihood estimator as follows (Lu et al. 2018):

\[
L(\theta) = -m(\text{Ln}(\sigma^2) + \text{Ln}(|R|))
\]  

(10)

In Equation (10), \(\text{Ln}\) represents the logarithmic operator. The correlation matrix, \(R\), is determined as below:

\[
R = \begin{bmatrix}
1 & r(X_1, X_2) & \cdots & r(X_1, X_m) \\
r(X_2, X_1) & 1 & \cdots & r(X_2, X_m) \\
\vdots & \vdots & \ddots & \vdots \\
r(X_m, X_1) & r(X_m, X_2) & \cdots & 1
\end{bmatrix}
\]  

(11)

in which \(r(X_p, X_q)\) is the covariance basis function. Commonly, the Gaussian function is utilized to define the kernel function in the modeling process (Sun et al. 2017). By applying the Gaussian function for the correlation function, \(r(X_p, X_q)\) is given as follows:

\[
r(X_p, X_q) = \exp\left[-\sum_{i=1}^{k} \theta^i (X_p^i - X_q^i)^2\right]
\]  

(12)

where \(\theta^i, i = 1, 2, ..., k\) (where \(i\)-th is the correlation parameter for \(k\) input variables). The parameters of Equation (7) using the optimum condition of Equation (10) are given as follows:

\[
\beta = (G^T R^{-1} G)^{-1} G^T R^{-1} \theta
\]  

(13)

By comparing the kriging model, the BF is affected by the parameters of the kernel function in the stochastic term. Therefore, in the regression analysis of problems with limited data, the BFs are effective tools to provide accurately predicted results. Commonly, the second-order polynomial function is used for the kriging model by the following relation:

\[
G(X)^T \beta = \beta_0 + \beta_1 x + \frac{\beta_2}{1 + \exp(-\kappa x)}
\]  

(15)

The logistic function presented in Eq. (15) with sigmoid form can predict a nonlinear function for which its flexibility is controlled by using the factor \(\kappa\), which is determined by trial and error in this current work. The presented functions in Equations (14) and (15) are respectively used in the kriging model and the kriging-based logistic models proposed in this study. As can be realized, the presented logistic model is as simple as the standard kriging model, and it has all the abilities of the standard kriging models in the training phase in achieving perfect simulations.

The models presented in Eqs. (14) and (15) can be used for predicting the data in the testing (validation) phase under the input data as below:

Step 1: Determine \(r\) using Eq. (9) for point \(X\).
Step 2: Compute stochastic term \(r(X)^T \gamma\) using the attained results from step 1.
Step 3: Compute \(G(X)^T \beta\) for the point \(X\) using Eqs. (13) and (14).
Step 4: Predict target data using Eq. (7).

As can be seen, the unknown coefficients and correlation parameters are computed in the regression process using the Eqs. (10) and (13) for the training data points. The correlation matrix, \(R\), applied in Eqs. (8) and (13) is computed for the training data while the \(r(X)\) is determined as presented in step 1 for the testing data points.

The prediction in the testing dataset of these models should be evaluated based on several statistical models for accuracy, tendency, and agreement. We believe that this model could provide the most accurate results for problems with limited data points in the training phase subject to the number of training data being more than \(2k+1\).

**Application of models**

**Feature selection process**

Besides the importance of methods used for achieving proper prediction/simulation results in data mining problems, feature selection—which is the process of choosing the best and most relevant input parameters—plays an essential role in constructing the optimum data mining model. In other words, in the feature selection process, one confines the original input vector to an optimal one by eliminating some irrelevant or non-effective input parameters, which leads to a faster training procedure and more accurate results. There are a variety of different approaches for applying the feature selection process,
including the wrapper methods (e.g., forward feature selection) (Meyer et al. 2018), filter methods (e.g., the mutual information approach) (Cameron et al. 2019), and embedded methods (Rodriguez-Galiano et al. 2018).

In this study, based on the wrapper methodology, we used the forward selection combined with the \( k \)-fold cross-validation method, where tenfold were randomly assigned for applying the technique. The result of the feature selection process implies that just three input parameters, i.e., influent flow, influent temperature, and influent arsenic values, effectively influence the target value (effluent arsenic). As could be expected, influent arsenic values greatly affect predicting effluent arsenic (see Table 1). The interesting point is that the applied feature selection method selected only flow rate and temperature among the other six independent variables (flow rate, BOD, TSS, pH, temperature, and turbidity). It is worth mentioning that the flow rate parameter is the only quantitative parameter among the available independent variables so that no other parameter could represent its impact on effluent arsenic. On the other hand, temperature, as a qualitative parameter, was chosen to represent the impact of other quantitative parameters, such as BOD, TSS, pH, and turbidity. It does not mean that the other parameters do not directly affect the effluent arsenic; however, it implies that in this specific case study, taking into account the other input parameters, such as pH, will not improve the predictive results of the applied models. In Fig. 3, the final results of the feature selection process, based on the calculated \( R^2 \) (%), are illustrated. The highest calculated value for \( R^2 \) is attributed to selecting the influent arsenic, flow rate, and temperature.

**Implementation of the models**

In this study, MATLAB programming language was used to create artificial intelligence models. After implementing the models, the data was split into two datasets, 75% for the training phase and 25% for the testing period. Also, based on the trial-and-error process, the best parameters of different models including the number of hidden layers, activation function type, and learning rate were gained to minimize the difference between real and estimated arsenic concentration values.

In addition, the data are predicted using the kriging model by DACE MATLAB toolbox (Lophaven et al. 2002), while the calibrating process of the developed kriging-based logistic map is coded by MATLAB software with an extra subroutine that is connected to DACE MATLAB toolbox for searching the nonlinearity degree of the input variables using a logistic map.

**Modeling scenarios**

This study deals with two types of input combinations for constructing the data mining models; in the first scenario, all of the seven available independent variables are considered as the input vector (see Table 1). On the other hand, the second scenario is established based upon the effective independent variables elicited from the feature selection process: flow rate, influent arsenic, and temperature. Differentiating between these two input scenarios helps to evaluate the effectiveness and significance of the feature selection process in the final performance of the applied models in predicting effluent arsenic.

**Evaluation of the models’ performance**

To evaluate the performance of the models applied, several statistical measures are used, including the root mean square error (RMSE) and mean absolute error (MAE), as absolute scale deviance measures; the mean absolute geometric error (MAGE) and geometric reliability index (GRI), as relative scale deviance measures; and the Pearson’s correlation coefficient (PCC) and Nash-Sutcliffe efficiency (NSE) coefficient, as similarity measures (Jachner et al. 2007). The formula for each of the above-mentioned measures and the possible ranges and ideal values are presented in Table 2.

In Table 2, \( y \) and \( \hat{y} \) denote respectively the observed and simulated/predicted values for the influent/effluent arsenic concentrations in the training/testing stages. \( \tilde{y} \) is...
the average of the arsenic values in the desired period, and \( N \) is the number of data, which stands for the number of observed values in the training or testing stages.

**Results**

In this section, the simulated and predicted outcomes of the applied models in the training and testing stages are presented for the two input combination scenarios separately.

**First scenario, all available input parameters**

As mentioned, the first scenario considers all of the available independent parameters for the construction of the kriging and machine learning models. Table 3 shows the overall performance of the compared machine learning models in the training stage. The GBRT model shows a better training performance compared to the MARS and ANN models.

Given the results in Table 4 for the testing phase, it can be clearly seen that the kriging-logistic method excelled in the absolute deviance and similarity measures at predicting effluent arsenic values in comparison to the other applied models.

**Table 2** Statistical measures used in this study for the evaluation of models applied

| Classification of measures | Statistical measure (dimension) | Formula | Range and ideal value(s) |
|---------------------------|---------------------------------|---------|-------------------------|
| Absolute scale deviance measures | RMSE (mg/m³) | \( \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \) | \([0, +\infty)\) | Ideal= 0 |
| MAE | \( \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \) | \([0, +\infty)\) | Ideal= 0 |
| Relative scale deviance measures | MAGE | \( \exp \left( \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{y_i}{\hat{y}_i} \right) \right) \) | \([0, +\infty)\) | Ideal= 1 |
| GRI | \( \frac{1+\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{y_i}{\hat{y}_i} \right)^2}}{1-\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{y_i}{\hat{y}_i} \right)^2}} \) | \([1, +\infty)\) | Ideal= 1 |
| Similarity measures | PCC | \( \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i) \hat{y}_i}{\sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \sqrt{\sum_{i=1}^{N} (\hat{y}_i)^2}}} \) | \([-1, +1]\) | Ideal= -1 or +1 |
| NSE | \( 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2} \) | \([-\infty, 1]\) | Ideal= 1 |

**Table 3** Machine learning models’ performance for the first input scenario in the training phase

| Models | Absolute deviance measures (ug/L) | Relative deviance measures | Similarity measures |
|--------|-----------------------------------|---------------------------|---------------------|
|        | RMSE | MAE | MAGE | GRI | PCC | NSE |
| MARS   | 0.099 | 0.237 | 0.997 | 1.131 | 0.957 | 0.719 |
| GBRT** | 0.087 | 0.198 | 1.002 | 1.103 | 0.972 | 0.783 |
| ANN    | 0.098 | 0.231 | 0.997 | 1.129 | 0.958 | 0.726 |

**Table 4** Models’ evaluation based on the statistical measures for the first input scenario in the testing phase

| Category | Models | Absolute deviance measures (ug/L) | Relative deviance measures | Similarity measures |
|----------|--------|-----------------------------------|---------------------------|---------------------|
|          |        | RMSE | MAE | MAGE | GRI | PCC | NSE |
| Machine learning | MARS | 0.176 | 0.120 | 1.010 | 1.167 | 0.885 | 0.751 |
| GBRT** | 0.170 | 0.120 | 1.016 | 1.159 | 0.900 | 0.770 |
| ANN | 0.176 | 0.121 | 1.004 | 1.171 | 0.867 | 0.749 |
| Kriging based | Kriging | 0.206 | 0.142 | 1.010 | 1.200 | 0.819 | 0.655 |
| Kriging-logistic* | 0.136 | 0.104 | 1.009 | 1.121 | 0.930 | 0.850 |

Best obtained values are identified bold-faced; *the best predictive model; **the best machine learning model.
models. However, all models acted similarly for the relative deviance measures (MAGE and GRI). Specifically, the kriging-logistic method performed better than the other models for all of the statistic measures except the MAGE measure (1.009 for the kriging-logistic vs. 1.004 for the ANN model). By and large, the kriging-logistic method surpassed the other models in the prediction process of effluent arsenic. The scatter plots of the MARS, GBRT, kriging, and kriging-logistic models for the predicted values are shown in Fig. 4. They compare the modeled effluent arsenic concentration with the measured samples by using the coefficient of determination ($R^2$). In these plots, by the aid of the lines of agreement (1:1 line) and 20% error lines, it is shown how close the measured data to the predicted values are. The $R^2$ for the kriging-logistic model is 0.86, which means the predicted results match the measured arsenic values to a fair extent.

In Fig. 5, the predicted outcomes of the models, as well as the measured arsenic concentrations, are illustrated using violin plots. Similar to boxplots, violin plots are used to represent the comparison of a variable distribution (here the effluent arsenic) across various data categories. The median, mean, interquartile range (midspread), and kernel density estimation (the body of the violin plot) are presented in these plots. Taking into account the above characteristics of the plots reveals that the kriging-logistic and ANN models performed better than the other models. Obviously, the kriging method produced the least good results, as shown by the deformed kernel density shape and the interquartile range height compared to the corresponding plot of the measured data.

**Second scenario, modeling based on the feature selection procedure**

Similar to the previous section, the performances of the ML models using the statistical measures are given in Table 5. As can be seen, the ANN model acted slightly better in the training phase according to the absolute deviance measures and similarity measures. However, this does not imply this model’s superiority, and one should refer to the predicted results in the testing set for attaining a judgment.

Based on the calculated measures for the predicted arsenic effluent values in Table 6, it is obvious that the ANN model’s performance is not as accurate as that of the GBRT’s. However, the highest NSE (= 0.896) and PCC (= 0.962) scores in Table 6 show that the kriging-logistic method tended to be the best-fitted model among the five at predicting effluent arsenic. Likewise, the kriging-logistic method provided the lowest values for the absolute deviance measures ($RMSE = 0.113 \text{ ug/L}$ and $MAE = 0.083 \text{ ug/L}$), which shows that the kriging-logistic is the most accurate model, followed by
the GBRT (RMSE = 0.164 ug/L and MAE = 0.113 ug/L). Even though the ANN gave the best score for the MAGE, the kriging-logistic model ranked second (MAGE = 1.007) for the MAGE and first for the GRI (= 1.099). Subsequently, the kriging-logistic model is the dominant best predictive model concerning the prediction of effluent arsenic in this study.

In order to assess the performance of the developed models graphically, scatter plots and violin plots of the predicted values versus the measured data are depicted, respectively, in Figs. 6 and 7. First, the predicted effluent arsenic concentrations by the MARS, GBRT, kriging, and kriging-logistic models are compared with the measured data in terms of scatter plots in Fig. 6. In addition to having a higher coefficient of determination ($R^2=0.92$), it can obviously be seen from the provided scatter plots that the kriging-logistic method predicted values are closer to the corresponding measured effluent arsenic data than those of the other applied models. Given the trend line equations in the scatter plots (assume that the linear equation is $y = ax + b$), the kriging-logistic method followed by the MARS model provided closer $a$ and $b$ coefficients to 1 and 0, respectively.

Table 5 Machine learning models’ performance for the second input scenario in the training phase

| Models  | Absolute deviance measures (ug/L) | Relative deviance measures | Similarity measures |
|---------|----------------------------------|----------------------------|---------------------|
|         | RMSE    | MAE    | MAGE   | GRI     | PCC    | NSE   |
| MARS    | 0.102   | 0.248  | 0.997  | 1.131   | 0.955  | 0.702 |
| GBRT    | 0.111   | 0.238  | 1.008  | 1.114   | 0.961  | 0.645 |
| ANN     | 0.092   | 0.215  | 0.997  | 1.125   | 0.963  | 0.754 |

Table 6 Models’ evaluation based on the statistical measures for the second input scenario in the testing phase

| Category     | Models          | Absolute deviance measures (ug/L) | Relative deviance measures | Similarity measures |
|--------------|-----------------|----------------------------------|----------------------------|---------------------|
|              | RMSE    | MAE    | MAGE   | GRI     | PCC    | NSE   |
| Machine learning | MARS      | 0.174  | 0.113  | 1.015   | 1.167  | 0.883 | 0.757 |
|               | GBRT**  | 0.164  | 0.113  | 1.019   | 1.157  | 0.914 | 0.787 |
|               | ANN     | 0.170  | 0.113  | 1.006   | 1.166  | 0.878 | 0.764 |
| Kriging based | Kriging  | 0.177  | 0.126  | 1.009   | 1.181  | 0.879 | 0.747 |
|               | Kriging-logistic* | **0.113** | **0.083** | 1.099   | **0.962** | **0.896** |

Best obtained values are identified bold-faced; *the best predictive model; **the best machine learning model
In Fig. 7, the predicted results versus the measured arsenic concentration are summarized graphically using violin plots. Similar to the results obtained for the first scenario, all of the applied predictive models acted similarly in presenting the interquartile range. Nevertheless, the ML models show better performance for the measures of central tendency, including the mean and median. According to the rotated kernel density plot, it can also be easily observed that the predictions obtained from the
kriging-logistic and ANN models are more or less the same as the rotated density plot of the measured data.

Discussion

In general, the ranking of the applied machine learning models is similar for both the first and second input scenarios, being overall the GBRT, then the MARS, and ANN. However, on the whole, the kriging-logistic method provided the best predictive results for both input scenarios. The simple kriging method, on the other hand, gave the weakest performance among all of the models. To seek more insight into the results of the two input scenarios, we have compared the performance of both scenarios and presented the performance enhancement of the predicted values in the second scenario compared to the first scenario in Table 7.

As can be seen from the results in Table 7, the statistical measures related to the predicted arsenic values have been improved noticeably for the kriging methods, e.g., an averaged RMSE increase of 15%. This issue emphasizes the importance of the feature selection procedure in improving the performance of the data-driven statistical-based models, such as kriging models.

A similar claim could be made for the upgrading in the performance of the applied machine learning models (with the exception of the MAGE measure). However, the magnitude of improvement is not as high (an average RMSE increase of 2.7%) as the results of the kriging-based methods. It can be concluded that because of their complex structure, ML models are better tools in capturing the nonlinear properties of the system problem based on even some ineffective/irrelevant independent input parameters. In other words, ML models are less sensitive to the input vector parameters than the kriging methods.

Due to the fact that the applications of ML methods in modeling WWTPs are rare in literature, it is highly recommended that more related studies be accomplished in the field of using ML methods in simulating WWTPs. This should improve the knowledge and understanding of ML methods’ ability to evaluate and predict the efficiency of WWTPs.

Conclusions

This study assessed and evaluated the potential of kriging-based and machine learning models in predicting the effluent arsenic concentration of a treatment plant using feature selection procedure establishing two different input scenarios. To meet this end, monthly influent wastewater discharge and six quality parameters, as well as effluent arsenic concentrations, were gathered from the Point Loma WWTP, which is a primary wastewater treatment facility in San Diego, California. The performance of both model categories was evaluated in terms of statistical measures, including absolute scale deviance measures (RMSE and MAE), relative scale deviance measures (MAGE and GRI), and similarity measures (PCC and NSE), as well as diagnostic methods such as scatter and violin plots. Results showed that ML models, such as ANN, MARS, and GBRT, could be effectively applied to the monthly interval prediction of arsenic concentration of effluent. Among the ML models, overall, the GBRT model showed a slightly higher prediction accuracy in the testing stage. The standard kriging model was unsuccessful in giving acceptable predictive results compared to the other models applied. However, it was revealed that the modified kriging model, namely the kriging-logistic model, provided the best predictive results in comparison to the ML models. It is worth mentioning that the kriging-based models were more sensitive to the outcomes of the feature selection procedure than the ML models. To sum up, this study suggested the efficiency and robustness of machine learning models and the higher accuracy of kriging-logistic method in modeling effluent arsenic of WWTPs.

Acknowledgements As the corresponding author, I wish to express my appreciation to the Alexander von Humboldt Foundation for providing financial support for this research project within the framework of the Return Fellowship program. Also, the third author was supported by University of Zabol under grant nos UOZ-GR-9018-1 and UOZ-GR-9719-1.

Author contribution Mohammad Zounemat-Kermani designed the study and carried out the analyses. Meysam Alizamir and Behrooz Keshtegar executed the models. Mohammad Zounemat-Kermani, Okke
Batelaan, and Reinhard Hinkelmann wrote the paper. All authors read and approved the final manuscript.

**Funding** This work was supported by the Alexander von Humboldt Foundation under the Georg Forster Research Fellowship program received by Mohammad Zounemat-Kermani as the applicant and Reinhard Hinkelmann as the host. Also, the third author was supported by University of Zabol under grant numbers UOZ-GR-9618-1 and UOZ-GR-9719-1.

**Data availability** The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

**Declarations**

**Ethics approval** The submitted study has not been published elsewhere in any form or language (partially or in full).

**Consent to participate and publish** All authors agreed with the content and all gave consent to submit and publish the research paper.

**Competing interests** The authors declare no competing interests.

**References**

Agarwal A, Upadhyay U, Sreedhar I, Singh SA, Patel CM (2020) A review on valorization of biomass in heavy metal removal from wastewater. J Water Process Eng 38:101602

Alizamir M, Kim S, Kisi O, Zounemat-Kermani M (2020a) A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: case studies of the USA and Turkey regions. Energy 197:117239

Alizamir M, Kim S, Zounemat-Kermani M, Heddam S, Kim NW, Singh VP (2020b) Kernel extreme learning machine: an efficient model for estimating daily dew point temperature using weather data. Water 12(9):2600

Alтоваtі WAH, Haris SA, Almoaіhаmі H, Shahіr S, Zakаrіa Z, Ibrahіm S (2020) The removal of arsenic species from aqueous solution by indigenous microbes: batch biosorption and artificial neural network model. Environ Technol Innov 19:100830

Ansari M, Othman F, Abunama T, El-Shafie A (2018) Analysing the accuracy of machine learning techniques to develop an integrated influent time series model: case study of a sewage treatment plant, Malaysia. Environ Sci Polit Res 25(12):12139–12149

Bekkari N, Zeddouri A (2019) Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. Management of Environmental Quality: An International Journal

Bernardelli A, Marsili-Libelli S, Manzini A, Stancari S, Tardini G, Montanari D et al (2020) Real-time model predictive control of a wastewater treatment plant based on machine learning. Water Sci Technol 81(11):2391–2400

Cameron TG, Jackel B, Oliveira DM (2019) Using mutual information to determine geoeffectiveness of solar wind phase fronts with different front orientations. J Geophys Res Space Physics 124(3):1582–1592

Fard MB, Mirbagheri SA, Pendashiteh A, Alavi J (2020) Estimation of effluent parameters of slaughterhouse wastewater treatment with artificial neural network and B-spline quasi interpolation. Int J Environ Res 14(5):527–539

Fei CW, Li H, Liu HT, Lu C, Keshtegarc, An LQ (2020) Multilevel nested reliability-based design optimization with hybrid intelligent regression for operating assembly relationship. Aerosp Sci Technol 103:105906

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 1–67

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Friedman JH, Roosen CB (1995) An introduction to multivariate adaptive regression splines. Stat Methods Med Res 4(3):197–217

Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14-15):2627–2636

Ge J, Guha B, Lippincott L, Cach S, Wei J, Su TL, Meng X (2020) Challenges of arsenic removal from municipal wastewater by coagulation with ferric chloride and alum. Sci Total Environ 725:138351

Giri AK, Patel RK, Mahapatra SS (2011) Artificial neural network (ANN) approach for modelling of arsenic (III) biosorption from aqueous solution by living cells of Bacillus cereus biomass. Chem Eng J 178:15–25

Granata F, Papirio S, Esposito G, Gargano R, De Marinis G (2017) Machine learning algorithms for the forecasting of wastewater quality indicators. Water 9(2):105

Guo H, Jeong K, Lim J, Jo J, Kim YM, Park JP et al (2015) Prediction of effluent concentration in a wastewater treatment plant using machine learning models. J Environ Sci 32:90–101

Haghiahi AH, Nasrolahi AH, Parsae A (2018) Water quality prediction using machine learning methods. Water Qual Res J 53(1):3–13

Harper TR, Kingham NW (1992) Removal of arsenic from wastewater using chemical precipitation methods. Water Environ Res 64(3):200–203

Hasanipanah M, Meng D, Keshtegarc, Trung NT, Thai DK (2021) Nonlinear models based on enhanced Kriging interpolation for prediction of rock joint shear strength. Neural Comput & Applic 33(9):4205–4215

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

Heddam S, Kisi O (2018) Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 559:499–509

Heddam S, Keshtegarc, Kisi O (2020) Predicting total dissolved gas concentration on a daily scale using kriging interpolation, response surface method and artificial neural network: Case study of Columbia river Basin Dams, USA. Nat Resour Sci 29(3):1801–1818

Huang D, Allen TT, Notz WI, Miller RA (2006) Sequential kriging optimization using multiple-fidelity evaluations. Struct Multidis—cip Optim 32(5):369–382

Hubadillah SK, Othman MHD, Ismail AF, Rahman MA, Jaafar J (2019) A low cost hydrophobic kaolin hollow fiber membrane (h-KHFM) for arsenic removal from aqueous solution via direct contact membrane distillation. Sep Purif Technol 214:31–39

Icke O, van Es DM, de Koning MF, Wuister JJG, Ng J, Phua KM et al (2020) Performance improvement of wastewater treatment processes by application of machine learning. Water Sci Technol 82(12):2671–2680

Jachner S, Van den Boogaard G, Potzold T (2007) Statistical methods for the qualitative assessment of dynamic models with time delay (R Package qualV). J Stat Softw 22(8):1–30

Jiang C, Qiu H, Gao L, Wang D, Yang Z, Chen L (2020) Real-time estimation error-guided active learning kriging method for time—dependent reliability analysis. Appl Math Model 77:82–98

Keshtegarc, Mert C, Kisi O (2018) Comparison of four heuristic—tic regression techniques in solar radiation modeling: kriging
method vs RSM, MARS and M5 model tree. Renew Sust Energ Rev 81:330–341
Kim S, Alizamir M, Zounemat-Kermani M, Kisi O, Singh VP (2020) Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. J Environ Manag 270:110834

Komorowicz I, Baralkiewicz D (2016) Determination of total arsenic and arsenic species in drinking water, surface water, wastewater, and snow from Wielkopolska, Kujawy-Pomerania, and Lower Silesia provinces, Poland. Environ Monit Assess 188(9):1–22

Krige DG (1952) A statistical approach to some basic mine valuation problems on the Witwatersrand, by DG Krige, published in the Journal, December 1951; interim reply by the author to the discussion. J South Afr Inst Min Metall 52(11):254–266

Lophaven SN, Nielsen HB, Søndergaard J (2002) DACE: a Matlab kriging toolbox (Vol. 2). IMM, Informatics and Mathematical Modelling, The Technical University of Denmark

Lu C, Feng YW, Lien RP, Fei CW (2018) Improved kriging with extremum response surface method for structural dynamic reliability and sensitivity analyses. Aerosp Sci Technol 76:164–175

Lu C, Feng YW, Fei CW, Bu SQ (2019) Improved decomposed-coordinated kriging modeling strategy for dynamic probabilistic analysis of multicomponent structures. IEEE Trans Reliab 69(2):440–457

Mandal S, Mahapatra SS, Patel RK (2015) Neuro fuzzy approach for arsenic (III) and chromium (VI) removal from water. J Water Process Eng 5:58–75

Matheron G (1963) Principles of geostatistics. Econ Geol 58(8):1246–1266

Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T (2018) Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. J Environ Manag 270:110834

Sheikhi S, Faraji Z, Aslani H (2021) Arsenic health risk assessment and the evaluation of groundwater quality using GWQI and multivariate statistical analysis in rural areas, Hashtroud, Iran. Environ Sci Pollut Res 28(3):3617–3631

Şimsek EB, Alkay T (2020) Artificial neural network approach for the prediction of effluents streams from a wastewater treatment plant: a case study in Kocaeli (Turkey). Mugla J Sci Technol 6(1):164–171

Song Y, Zhou H, Wang P, Yang M (2019) Prediction of clathrate hydrate phase equilibria using gradient boosted regression trees and deep neural networks. J Chem Thermodyn 135:86–96

Sun Z, Wang J, Li R, Tong C (2017) LIF: A new Kriging based learning function and its application to structural reliability analysis. Reliab Eng Syst Saf 157:152–165

Wang A, Zhou K, Zhang X, Zhou D, Peng C, Chen W (2020) Arsenic removal from highly-acidic wastewater with high arsenic content by copper-chloride synergic reduction. Chemosphere 238:124675

Xiao M, Zhang J, Gao L, Lee S, Eshghi AT (2019) An efficient Kriging-based subset simulation method for hybrid reliability analysis under random and interval variables with small failure probability. Struct Multidiscip Optim 59(6):2077–2092

Xiao M, Zhang J, Gao L (2020) A system active learning Kriging method for system reliability-based design optimization with a multiple response model. Reliab Eng Syst Saf 199:106935

Yapıcıoğlu P (2018a) Investigation of environmental-friendly technology for a paint industry wastewater plant in Turkey. Suleyman Demirel Universitasi Fen Bilimleri Enstitüsü Dergisi 22(1):98–106

Yapıcıoğlu PS (2018b) Environmental impact assessment for a meat processing industry in Turkey: wastewater treatment plant. Water Pract Technol 13(3):692–704

Yapıcıoğlu P (2020) Grey water footprint assessment for a dye industry wastewater treatment plant using Monte Carlo simulation: influence of reuse on minimisation of the GWF. International Journal of Global Warming 21(2):199–213

Yetis AD, Yesilnacar MI, Atas M (2021) A machine learning approach to dental fluorosis classification. Arab J Geosci 14(2):1–12

Yun W, Lu Z, Jiang X (2018) An efficient reliability analysis method combining adaptive Kriging and modified importance sampling for small failure probability. Struct Multidiscip Optim 58(4):1383–1393

Zhang W, Goh AT (2016) Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci Front 7(1):45–52

Zhang M, Li Y, Long X, Chong Y, Yu G, He Z (2018) An alternative approach for nitrate and arsenic removal from wastewater via a nitrate-dependent ferrous oxidation process. J Environ Manag 220:246–252

Zhang J, Gao L, Xiao M (2020) A new hybrid reliability-based design optimization method under random and interval uncertainties. Int J Numer Methods Eng 121(19):4435–4457

Zhou Z, Liu YG, Liu SB, Liu HY, Zeng GM, Tan XF et al (2017) Sorption performance and mechanisms of arsenic (V) removal by magnetic gelatin-modified biochar. Chem Eng J 314:223–231

Zounemat-Kermani M, Beheshti AA, Atae-Ashtiani B, Sabbagh-Yazdi SR (2009) Estimation of current-induced scour depth around pile groups using neural network and adaptive neuro-fuzzy inference system. Appl Soft Comput 9(2):746–755

Zounemat-Kermani M, Stephan D, Hinkelmann R (2019) Multivariate NARX neural network in prediction gaseous emissions within the influent chamber of wastewater treatment plants. Atmos Pollut Res 10(6):1812–1822
