Earth Science Informatics
https://doi.org/10.1007/s12145-022-00828-1

RESEARCH ARTICLE

Salinity-constituent conversion in South Sacramento-San Joaquin Delta of California via machine learning

Peyman Namadi[1] · Minxue He[1] · Prabhjot Sandhu[1]

Received: 15 February 2022 / Accepted: 25 May 2022

Abstract

The levels of total salinity and its ion constituents in estuarine environments are important indicators of the overall suitability of water for environmental, agricultural, and urban use. These constituents include Total Dissolved Solids (TDS), dissolved chloride (Cl^-), dissolved sulfate ($SO4^{2-}$), dissolved sodium (Na^+), dissolved calcium (Ca^{2+}), dissolved magnesium (Mg^{2+}), dissolved nitrate (NO3), dissolved potassium (K), dissolved bromide (Br^-), dissolved boron (B), Alkalinity, and hardness, among others. In practice, salinity is typically measured indirectly as electrical conductance (EC) via automatic sensors while the concentration of each constituent is often measured from discrete water samples (i.e., grab samples) and thus is available much less frequently than salinity. Quadratic regression equations are generally developed between salinity (as the predictor, represented by EC) and individual constituents (as the predictand) based on grab sample data. The regression models are then applied to estimate the concentrations of individual constituents given EC during the period when grab samples are not available. The current study develops four types of machine learning models: the Generalized Additive Model, Regression Trees, Random Forest, and Artificial Neural Networks, to emulate conventional regression equations in salinity-constituent conversion. A case study in the South Sacramento-San Joaquin Delta of California is presented to illustrate the development and application of these models and to compare their performance with that of the benchmark regression models. The results indicate that machine learning models can provide comparable or superior simulations to the regression models. Among the four machine models examined, the Random Forest model tends to yield the best results. The study further discusses the scientific and practical implications of the models proposed, their limitations, as well as future work to further improve their performance and applicability.

Keywords Salinity · Ion constituents · Sacramento-San Joaquin Delta · Machine learning · Random Forest

Introduction

Background

The overall quality of water, including the levels of various ion constituents, determines its suitability for environmental, agricultural, and urban use (WHO 2008; Li and Wu 2019). Especially, human health is directly affected by the quality of drinking water, and access to safe drinking water is a basic human right for all people worldwide (WHO 2017). Water quality varies spatially and temporally, depending on the source water conditions. This is particularly evident in regions with multiple water sources of contrasting levels of constituents, including the Sacramento-San Joaquin Delta (Delta) in California, United States (U.S.). The Delta serves as the hub of California's major water distribution systems. It receives seasonal freshwater inflows from the watersheds on the east and salty tides from the Pacific Ocean on the west (Fig. 1). Water is pumped from the Delta to its surrounding regions as well as the drier Southern half of California to support over two-thirds of the State's population and over 15,000 km^2 of farmlands via the State Water Project (SWP), the Central Valley Project (CVP), and local water distribution systems (Becker et al. 1976; Sabet and Coe 1986; He et al. 2020).

Environment and ecosystem protection in the Delta is critical to SWP and CVP operations and management. Delta

Communicated by: H. Babaie

✉ Minxue He
kevin.he@water.ca.gov

[1] Delta Modeling Section, Department of Water Resources, Bay-Delta Office, Sacramento, CA, USA

Published online: 28 June 2022

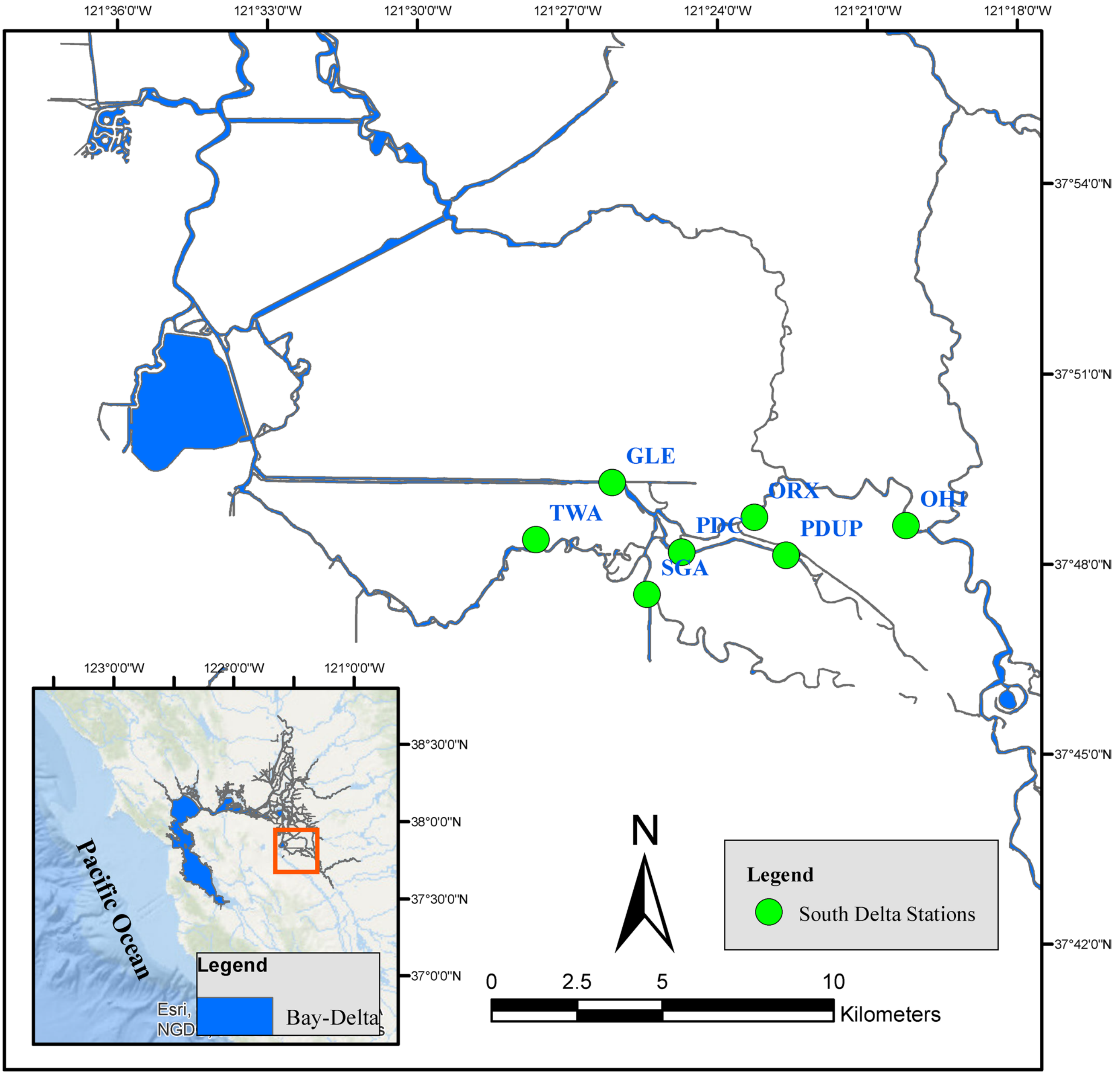


Fig. 1 Map of stations with continuous water quality and ion concentration measurements in the South Delta. The insert map shows the location of the San Francisco Bay and Sacramento-San Joaquin Delta (Bay-Delta) which contains the study area South Delta (highlighted in the red rectangle)

is a habitat of about 750 species of plants and animals of which some are endangered or threatened species (Healey et al. 2016). For instance, Delta smelt, a small and relatively obscure native fish that can only tolerate low levels of salinity, was listed as threatened under the U.S. Federal Endangered Species and California Endangered Species Act. Changes in Delta water quality is the biggest cause of Delta smelt population decline (Bennett 2005; Moyle et al. 2018). Water quality standards centered on the total salinity level as well as dissolved chloride (Cl^-) have been developed to ensure that the water across the Delta, particularly at the intakes of SWP and CVP, is appropriate for environmental, drinking, agricultural, and other purposes (CSWRCB 1999; USFWS 2008). Additionally, other constituents such as Total Dissolved Solids (TDS), dissolved sulfate ($SO4^{2-}$), dissolved sodium (Na), dissolved calcium (Ca), dissolved magnesium (Mg^{2+}), dissolved nitrate (NO3), dissolved potassium (K), dissolved bromide (Br^-), dissolved boron (B), Alkalinity, and hardness are also important for water resources management in the Delta.

The cause of salty taste in the drinking water is high level (> 250 mg/l) of chloride (Kumar and Puri 2012). Bromide is important because of its role in organic carbon and drinking water disinfectants to form disinfection byproducts such as bromated, brominated trihalomethanes, and brominated haloacetic acids. These disinfection byproducts are suspected carcinogens, and their concentrations are regulated by the U.S. Environmental Protection Agency (EPA). The maximum threshold for sulfate level in the water is 250 mg/l based on aesthetic effects (i.e., taste and odor) (Lopez et al. 2017). Consuming too much nitrate can affect how blood carries oxygen and can cause methemoglobinemia, also known as a blue baby syndrome (Lewis and Donald 1986). Concentrations of boron in water vary largely, depending on the drinking-water sources. In most cases, the variations range from 0.1 to 0.3 mg/l. In general, 1.5 and 2 mg of total daily intake boron is the maximum safe concentration for human consummation (WHO 2009). Based on WHO standards, sodium (Na^+) level in the drinking water should be less than 200 mg/l. High level of sodium can cause many fatal diseases such as kidney damage and hypertension (Mohsin et al. 2013). Also, WHO recommends 75 and 150 mg/l as the threshold levels for Ca and Mg^{2+}, respectively (Mohsin et al. 2013).

Literature review

Salinity increases the ability of a solution to conduct an electrical current. Therefore, the salinity level highly correlates with the Electrical Conductance (EC) of the water. In the Delta, salinity is generally measured by EC which is typically monitored via automatic sensors. In comparison, other ion constituents are normally not monitored directly by sensors. Instead, discrete grab samples are collected and analyzed for different types of constituents. The common practice is to develop regression models that map different constituents to EC, which is readily available. Guivetchi (1986) analyzed the sample data collected from Delta on Total Dissolved Solids and chloride concentration. He developed linear regression equations between them and EC, respectively. Specifically, he categorized the data according to the type of water year when the data were collected. He developed linear regression equations for different water year types (dry, normal, wet) and all years, respectively, at selected locations in the Delta. Although these linear models were very beneficial, water quality in some areas such as the south and central Delta depends on diverse water sources, which vary not only across different types of water years but also across different seasons (e.g., wet season, dry season) in a single year. A single linear equation, classified just in terms of water year type, could not learn the behavior of the seasonal changes. The California Urban Water Agencies (CUWA 1995) found that bromide, sulfate, total dissolved solids, and chloride grab sample data were limited by the regression relationships for the two main origins: agricultural drainage and seawater.

Suits (2002) divided water quality stations within the Delta based on the resemblance of their site-specific relationship between chloride and calcium. The calcium to chloride ratio is the lowest in the West Delta and the highest where freshwater dominates (Sacramento River area; the northeast portion of the Bay-Delta illustrated in the insert map of Fig. 1). Hutton (2006) studied the relationship between chloride and sulfate ions. He identified that their ratio varies from about 7 where seawater dominates down to about 1 where San Joaquin River (the southeast portion of the Bay-Delta illustrated in the insert map of Fig. 1) water dominates. He also provided diagrams of different mineral ions as a function of EC and Alkalinity but did not quantify any of these correlations. However, he showed how these EC relationships varied depending on the modeled origin water contributions based on simulations from a Delta hydrodynamic and water quality model named Delta Simulation Model II (DSM2). Liu and Suits (CDWR 2012) assessed the performance of different methods in simulating bromide at select spots in the Delta, i.e., simulating bromide with DSM2 or converting EC to bromide via regression equations based on: (a) in situ samples; (b) volumetric fingerprints; and (c) EC-fingerprint-based multiple regressions. They concluded that Delta-wide regression based on DSM2-simulated EC functioned as good as direct bromide simulation using DSM2. They also concluded that, although site-specific regressions performed the best, regional regressions performed fairly close to site-specific regressions.

Denton (2015) used grab sample data to develop non-linear relationships for key water quality constituents of concern as a function of EC. The work focused on chloride, bromide, sodium, calcium, sulfate, magnesium, potassium, Alkalinity, and hardness. Corresponding relationships between these constituents and TDS were also developed. Specifically, Denton (2015) developed separate non-linear quadratic models for three boundary conditions in Delta: seawater-dominated (western Delta), freshwater-dominated (Sacramento River), and agricultural drainage-dominated (San Joaquin River).

Chen and Roy (2015) developed multi-layer perceptron (MLP) neural networks to estimate the volumetric contributions of individual flows in the Delta. Results of DSM2 were used to train those MLP models. Model inputs were similar to that of DSM2 (i.e., eight flow boundaries and six temporary barrier boundaries). Their results showed that MLP models could emulate DSM2 with acceptable performance while processing time and usability were the concerns.

Most recently, the California Department of Water Resources (CDWR) collected ion samples from at a number of stations in South Delta (NCRO 2021). Quadratic equations were developed to relate 12 ion constituents to EC, following the approach of Denton (2015). They found that these quadratic equations worked fairly well in most cases but performed poorly for a few constituents.

Scope of the current work

In spite of those extensive efforts in modeling ion constituents from salinity in the Delta, there is currently no effective model to simulate the concentrations of all ion constituents in the Delta with desirable performance. Specifically, most of these previous studies focused on a subset of ion constituents. Regression models developed for some constituents (e.g., nitrate, boron, and potassium) have fairly poor performance. The current study aims to bridge that gap by developing machine learning models to emulate the conventional regression equations in simulating the concentrations of these 12 ion constituents given salinity and auxiliary information in South Delta. To that end, four types of widely used supervised machine learning models: the Generalized Additive Model (GAM), Regression Trees (RT), Random Forest (RF), and Artificial Neural Networks (ANNs) are explored. The most recent grab samples collected in South Delta (NCRO, 2021) are employed to exemplify the feasibility and applicability of these proposed models. The scientific and practical values of the study will be discussed, along with the limitations and future work.

Materials and methods

Study locations and study dataset

The California Department of Water Resources (CDWR) collected standard ion samples at seven representative stations (Fig. 1) co-located with water quality sensors that continuously measure salinity conditions (reported as EC) in South Delta from 2018 to 2020. Samples were collected roughly on a monthly basis for ion analysis at 1-meter depth using a Van Dorn sampler at these locations. For a particular grab sample, the concentrations of the aforementioned 12 constituents were determined via laboratory analysis (NCRO 2021).

After data quality control, a total number of 183 samples were identified and used to train and test the proposed machine learning models. Figure 2 shows the scatter plots of observed salinity (represented by EC) at study locations versus the concentrations of those 12 ion constituents. In these plots, data collected from different stations are distinguished via different colors. The figure shows the relationship between EC (x-axis) and ion constituents (y-axis) for all samples. Except for nitrate (Fig. 2g), potassium (Fig. 2h), and boron (Fig. 2j), other constituents exhibit a fairly evident monotonic relationship with EC. The concentrations of these nine constituents generally increase with increasing EC. Among all seven locations, PDUP (purple dots in Fig. 2) has the highest salinity level and highest concentrations of these nine constituents. In comparison, PDUP has the lowest nitrate level (Fig. 2g), moderate boron level (Fig. 2j), and moderate to high level of potassium (Fig. 2h) among all stations. Excluding PDUP, potassium and boron at the remaining six locations tend to vary monotonically with EC. However, this is not the case for nitrate of which the variation pattern with EC is not visually distinct even with PDUP samples excluded.

Figure 3 illustrates the Pearson correlation coefficients between each pair of the 13 study variables which consists of EC and 12 ion constituents. It is clear that, except for nitrate, potassium, and boron, the other nine constituents have an extremely strong linear correlation with EC (correlation coefficients over 0.94). These nine constituents are also strongly correlated with each other, though their concentration levels vary widely (vertical axis of Fig. 2). In contrast, nitrate has very weak linear correlations with all other variables. The absolute correlation coefficient is generally less than 0.3. Potassium and boron have moderate correlations with EC. The associated correlation coefficients are 0.69 and 0.73, respectively. Their correlations with other variables except for EC and nitrate are also moderate. The observations in Figs. 2 and 3 suggest that regression equations may not be sufficient to accurately portray the relationships between salinity and a number of ion constituents, particularly nitrate, potassium, and boron.

Proposed models, study scenarios, and study metrics

Four supervised machine learning (ML) techniques: the Generalized Additive Model (GAM), Regression Trees (RT), Random Forest (RF), and Artificial Neural Networks (ANNs), were employed to estimate ion constituents given EC at these seven study stations. Due to different water quality behaviors in each station, in the first scenario (Table 1), station name was employed as a categorical variable (along with salinity measurements at each station) as the input features to the ML models. Considering that machine learning algorithms cannot use categorical variables in the numerical calculation, the encoding technique was implemented to convert the seven stations' names to numerical values. In encoding the categorical variables, a number (1 to 7) is assigned to each station. The numbers have no quantitative value, and the order does not matter (Potdar et al. 2017). Based on the results from the first scenario, the ML model with the most desirable performance was selected and further explored in the second scenario (Table 1). In this scenario, the month and the type of the water year (He et al. 2021) (when a specific sample was taken) were added as additional input features to assess their potential impacts on the model outcome. In the first scenario, ML models were trained for the entire dataset (100% of the dataset) to maintain consistency

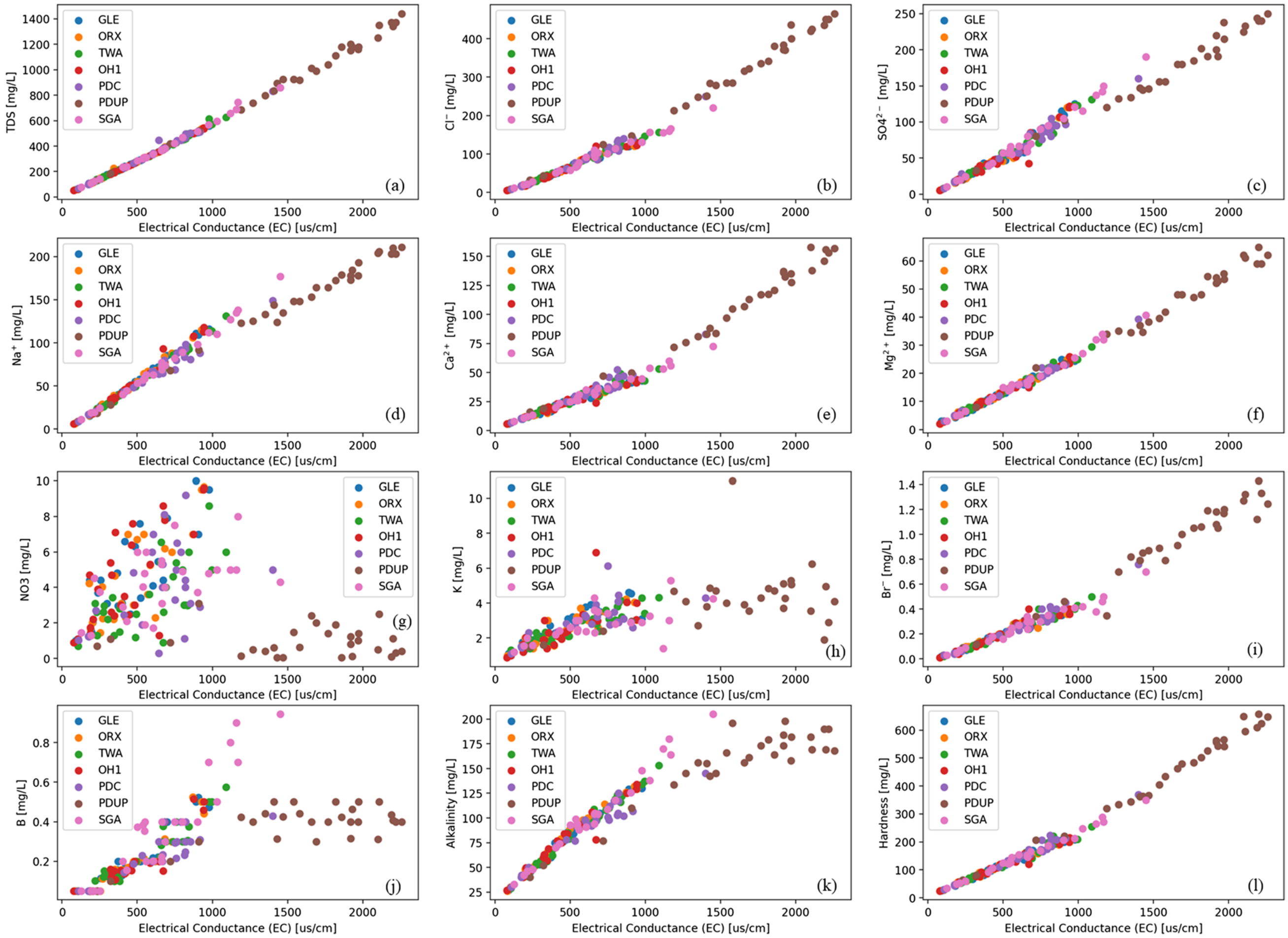


Fig. 2 Scatterplot matrix showing the relationship between Electrical Conductance (EC) and 12 ion constituents at study locations. Panels (**a**), (**b**), (**c**), (**d**), (**e**), (**f**), (**g**), (**h**), (**i**), (**j**), (**k**), and (**l**) are scatter plots that show EC against TDS, Cl^-, $SO4^{2-}$, Na^+, Ca^{2+}, Mg^{2+}, NO3, K, Br^-, B, Alkalinity, and hardness, respectively. Note that the scales on y-axes of different panels are different

with the training method applied in developing the quadratic regression models. The input-output datasets were randomly split into two groups for training (80% of the dataset) and testing (20% of the dataset) in the second scenario.

The performance of four ML models and the quadratic equations (as the benchmark model) was evaluated using a total number of six criteria consisting of correlation coefficient (R), coefficient of determination (R^2), Mean Absolute Error (MAE), Standard Deviation (SD), centered Root Mean Squared Difference (RMSD), and percent Bias (Table 2). R^2 ranges from 0 to 1, with a value close to 1 meaning that model simulations capture most of the variability in the observed data. MAE and RMSD are positive numbers, with a value close to 0 meaning that the modeled values are very close to observed values. Percent bias shows whether the model over-estimates or under-estimates the target variables and by how much on average. SD measures the overall dispersion of the observed or simulated variables. A brief overview of the supervised machine learning techniques proposed in this study is provided as follows.

Generalized Additive Model (GAM)

The Generalized Additive Model (GAM) provides a general framework for improving standard linear models by allowing for non-linear relationships between each feature and the response (Hastie and Tibshirani 1986; James et al. 2013). GAM replaces each linear component with a (smooth) non-linear function $f_j(x_{ij})$ and calculates a separate $f_j(x_{ij})$ for each predictor when others remain fixed. GAM divides the variation range of each environmental predictor (EC and station name in this case) into distinct regions. It fits a polynomial function in each region with the limitation that the polynomial function in each region needs to join smoothly to the polynomial in the next region or zone. Equation 1 shows the general form of the GAM model.

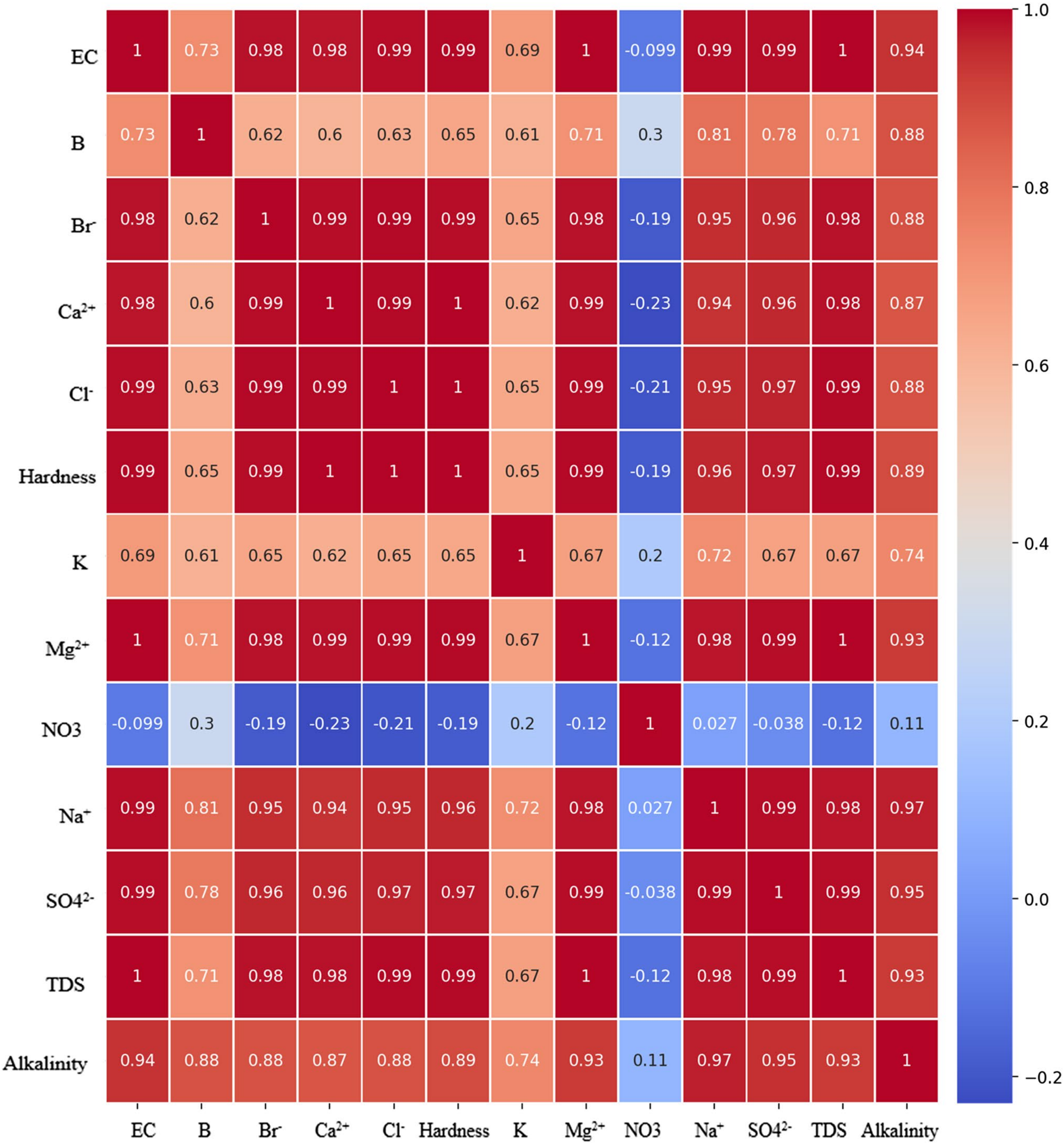


Fig. 3 Pearson correlation coefficients among 13 study variables consisting of salinity (represented by EC) and the 12 ion constituents

Table 1 Machine learning modeling scenarios

Scenario	Scenario 1	Scenario 2
Name	RF-2	RF-4
ML Models	GAM, RT, RF, ANNs	Best ML Model
Predictors	EC Station Name	EC Station name Month Water year type

$$y_i = \beta_0 + f_1(EC) + f_2(Station\ Name) + \epsilon_i,\ (i = 1,\ \ldots,\ n) \quad (1)$$

where y_i represents the targets that are ion constituents in our study. Also, $f_1(EC)$ is unspecified smooth ("nonparametric") function of EC. β_0 is the intercept, ϵ_i is the error variable, and n is number of samples that is 183 in our study. The individual functions of the GAM model were developed using

Table 2 Study metrics

Abbreviation	Description	Formula
R	Pearson Correlation Coefficient	$\frac{\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{\widehat{y}}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{\widehat{y}}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$
R^2	Coefficient of Determination	$1-\frac{\mathrm{SSE}}{\mathrm{SSTotal}}$
MAE	Mean Absolute Error	$MAE=\frac{\sum_{i=1}^{n}\lvert\widehat{y}_i-y_i\rvert}{\mathrm{n}}$
SD	Standard Deviation	$\sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i-\bar{\widehat{y}})}{n-1}}$
RMSD	Centered Root Mean Squared Difference	$\sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{\widehat{y}}\right)-\left(y_i-\bar{y}\right)}{n}}$
Bias	Percent Bias	$\frac{\sum_{i=1}^{n}(\widehat{y}_i-y_i)}{\sum_{i=1}^{n}y_i}\times 100$

SSE: Sum of squared error (or residuals).$SSE=\sum_i(y_i-\widehat{y}_i)^2$

SSTotal: Sum of squared deviations from the mean $\underline{y}$).SSTotal $=\sum_i(y_i-\widehat{y}_i)^2+\sum_i(\widehat{y}_i-\bar{y})^2$

$y_i=$ *observed values*

$\underline{\widehat{y}_i}=$ *predicted values*

$\underline{y}$= *mean of observed values*

$\widehat{y}$=*mean of predicted values*

the *mgcv* estimation package (Wood 2017) in the R statistical computing environment (R Core Team 2021).

Regression Trees (RT)

Classification and Regression Trees (CART) are popular machine learning methods that can be applied to both regression (Regression Trees) and classification (Classification Trees) problems. This method stratifies the predictor space into several regions and assigns a mean of each region to all observed data included in a specific rectangular region (Breiman et al. 1984; Loh 2011; James et al. 2013). Tree-based ML models are useful for interpretation, as their results indicate the importance of predictors and the split points suggest the best threshold for each predictor.

The first step in each decision tree is finding the best split predictor and cutpoint at each node of the decision tree. The model implements the recursive binary splitting method that splits the dataset into two new branches. The decision tree considers all predictors and all possible cutpoints for each predictor and then chooses the predictor and cutpoints of which the Residual Sum of Squares (RSS) is the minimum (James et al. 2013). Equation 2 shows the RSS criteria that should be minimized at each splitting point, where R_1 and R_2 are the two new branch regions after each splitting process, j is the predictor indicator, and s is the cutpoint. y_i represents the targets that are ion constituents in our study, and x_i represents the predictors. The individual functions of the RT model were determined by using the *rpart* package (Therneau and Atkinson 2019) in the R statistical computing environment.

$$RSS=\sum_{i:x_i\in R_1(j,s)}\left(y_i-\widehat{y}_{R_1}\right)^2+\sum_{i:x_i\in R_2(j,s)}\left(y_i-\widehat{y}_{R_2}\right)^2 \tag{2}$$

Random Forest (RF)

Random Forest (RF) has demonstrated strong predictive performance in addressing a wide range of classification and regression analysis problems (Ho 1995; Breiman 1999; Liaw and Wiener 2002; Namadi and Deng 2021). It incorporates multiple decision trees in conjunction with the bootstrap technique to decrease the variance of a statistical learning method. This allows for the production of new populations from the primary population by resampling data (James et al. 2013).

Theoretically speaking, the high variance of the regression tree model causes the overfitting problem, and it is important to limit the variance of regression tree models. Regression trees can learn complex relations in the data, but when trees grow deep, the final model has low bias and high variance that cause an overfitting problem. Furthermore, because of Gini impurity restrictions of regression trees, there is no guarantee that it uses all features. In contrast, Random Forest models rely on an ensemble of models (multiple decision trees). Random Forest is a substantial modification of bagging that builds a large collection of de-correlated trees. In fact, being able to choose these random subsets of features allows us to explore many different aspects of the entire feature space. Random Forest creates subsets of randomly picked features at each potential split. In light of that, the developer of the RF algorithm claimed that “Random forest models do not overfit” (Breiman 2001). Hastie et al. (2009) further ascertained that increasing the number of trees in RF does not cause the RF sequence to overfit since after a certain number of trees: (a) different random selections don’t reveal any more information, and (b) different random selections are simply duplicating trees that have already been created. Therefore, in theory, overfitting in the RF model is of minimal concern.

RF cumulates the results of all decision trees which were produced via bootstrapping. For instance, if β separate training datasets were produced by the bootstrapping method, $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^\beta(x)$ will be the results of each decision tree. Where $\hat{f}(x)$ is result of a regression tree and x is the model predictors such as EC and station name. Equation 3 shows the outcome of the RF model which is the average of all decision trees in order to obtain a single low-variance statistical learning model with more accuracy. The individual functions of the Random Forest model were determined by using the " RandomForest” package in the R statistical computing environment.

$$\hat{f}_{avg}(x) = \frac{1}{\beta}\sum\nolimits_{b=1}^{\beta}\hat{f}^b(x) \tag{3}$$

Artificial Neural Network (ANN)

Artificial intelligence-based neural network (ANN) models are predictive models that have been widely adopted for model identification, analysis, and forecasting. The ANN has been proven to be an effective method for developing non-linear relationships between a dependent variable and independent variables (McCulloch and Pitts 1943; Hopfield 1988; Zhang et al. 2015).

A typical ANN model consists of three primary layers: an input layer, a hidden layer, and an output layer. In this study, ANN consists of four layers: an input layer, two hidden layers, and an output layer. The input layer contains two input variables, namely, EC and station name as a categorical variable. The number of neurons in hidden layers and their activation functions were determined after experimenting multiple iterations until maximum simulation accuracy can be obtained. The number of neurons in each hidden layer was determined as 20 and the Rectified Linear Unit (ReLU) function $f(\alpha) = \max(o, \alpha)$ was selected as the activation function that returns 0 if it receives any negative input and returns α when the input value is positive. The loss function was the Mean Squared Error (MSE). Figure 4 shows the Artificial Neural Network architecture. The individual functions of the ANN model were determined by using the open source " H2O” package in the R statistical computing environment (Candel et al. 2016).

Results and discussion

This section first presents benchmark model (i.e., quadratic equation) results. Next, the section illustrates the performance of proposed ML models on simulating three ion constituents (nitrate, potassium, and boron), on which the benchmark models have a relatively poor performance. Next, the section focuses on the selection and evaluation of the ML model with the most desirable performance. Lastly, the section illustrates the performance of the selected model to the remaining nine ion constituents.

Benchmark model results

Using the same study dataset, CDWR (NCRO 2021) developed a quadratic equation for each ion constituent with the following form:

$$IC_i = C_{1,i} \times EC^2 + C_{2,i} \times EC + C_{3,i} \tag{4}$$

where IC_i is the concentration of the i-th ($i = 1, 2, \ldots, 12$) constituent; EC represents the salinity; C_1, C_2 and C_3 are coefficients. The coefficients of those quadratic equations are tabulated in Table 3.

Five metrics consisting of R^2, MAE, SD, RMSD, and percent Bias are calculated between regression model-generated simulations and their corresponding observations (Table 3). Based on R^2 values that represent the proportion of the

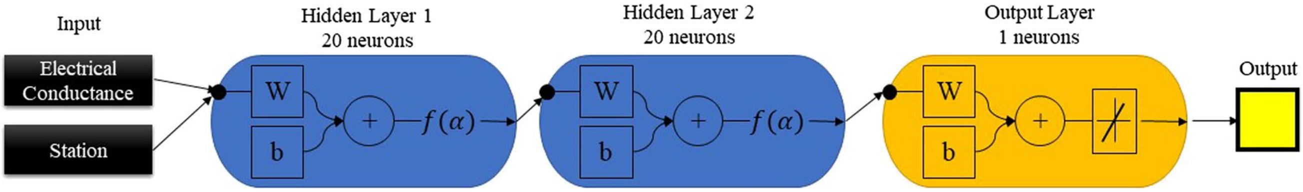


Fig. 4 Artificial Neural Network architecture

Table 3 Benchmark models (quadratic equations in the form of $IC = C_1 \times EC^2 + C_2 \times EC + C_3$) and their performance measured via statistical metrics

Ion Constituents	R^2	Bias (%)	MAE (mg/l)	SD (mg/l)	RMSD (mg/l)	C_1	C_2	C_3
B	0.75	1.7	0.06	0.15	0.09	-2.52E-07	7.74E-04	-0.1
Br^-	0.98	1.4	0.03	0.31	0.05	1.12E-07	3.68E-04	-0.02
Ca^{2+}	0.99	0.4	2.71	34.66	3.85	1.76E-05	2.99E-02	4.96
Cl^-	0.99	-0.1	6.25	105.75	9.7	4.14E-05	1.18E-01	-3.77
Hardness	0.99	0.2	8.13	145.05	11.74	5.07E-05	1.75E-01	14.25
K	0.59	-3.3	0.52	0.94	0.83	-1.35E-06	4.44E-03	0.89
Mg^{2+}	0.99	0.1	0.75	14.22	1.08	1.48E-06	2.46E-02	0.35
NO3	0.32	0.6	1.59	1.39	2.02	-4.28E-06	8.48E-03	0.93
Na^+	0.99	0.3	4.08	49.93	6.13	-1.84E-05	1.35E-01	-7.76
$SO4^{2-}$	0.98	0.2	5.76	57.37	8.04	-5.89E-06	1.24E-01	-6.73
TDS	0.99	-0.1	8.02	315.72	14.55	3.72E-05	5.37E-01	4.99
Alkalinity	0.96	0.2	5.37	42.4	8.63	-3.83E-05	1.59E-01	12.56

* Benchmark model coefficients adapted from (NCRO 2021)

variance for simulated and observed values, three ion constituents (NO3, K, and B) were not satisfactorily simulated. The R^2 values are 0.75, 0.59, and 0.32 for B, K, and NO3, respectively. For other nine constituents, the R^2 values are consistently higher than 0.96. Nevertheless, there are notable differences between model simulations and the observations on these nine constituents when measured by MAE (ranging from 0.03 to 8.02 mg/l) and RMSD (ranging from 0.05 to 14.55 mg/l). SD also varies largely across different constituents from 0.15 mg/l (B) to 315.72 mg/l (TDS). Percent bias which measures the average tendency of the simulated values to be larger or smaller than their observed ones shows that the quadratic models underestimate the concentrations of Cl^-, K, and TDS while overestimate the concentrations of the remaining constituents. However, the absolute bias values are generally small. All in all, the benchmark models perform well for all constituents but NO3, K, and B. The following three sub-sections presents the performance of proposed ML models on these three constituents, respectively.

Simulation of nitrate (NO3)

The performance of ML models on simulating the concentration of nitrate is first evaluated using two metrics, R^2 and Mean Absolute Error (MAE). The R^2 values are calculated as 0.32, 0.51, 0.67,0.88, 0.57 for the quadratic equation (benchmark), GAM, RT, RF, and ANN, respectively. The MAE values are determined as 1.62, 1.37, 1.09, 0.66, and 1.2 mg/l for these five models, respectively. The Random Forest (RF) model shows the best performance in simulating the concentration of NO3, yielding the highest R^2 and smallest MAE. Comparing the benchmark model (quadratic equation) to the RF model, the latter increases R^2 by 175% and decreases MAE by 59%.

Model performance is further illustrated via the Taylor diagram (Taylor 2001). Taylor diagrams illustrate the performance of different models by comparing the observed and estimated values through visualizing a series of points on a polar plot. The SD of simulated concentrations of nitrate from five models (benchmark and four ML models proposed) as well as their correlations with the observed nitrate and their RMSD from the corresponding observations are calculated and illustrated in Fig. 5. It is clear that, among five models, the outcome of the RF model (blue diamond in Fig. 5) has the highest correlation, lowest RMSD, and the closest SD to that of the observations. In comparison, the benchmark quadratic equation (red square) has the lowest correlation coefficient, highest RMSD, and the most different SD from that of the observations. In all four ML models, the performance of the RT model is generally inferior to that of the RF but better than that of the ANN model. The GAM model tends to have the lowest correlation coefficient, highest RMSD, and least desirable SD among four ML models.

Simulation of potassium (K)

The quadratic equation has a fairly poor performance in simulating the concentration of potassium. The corresponding R^2 is 0.59. In comparison, the R^2 values are 0.61, 0.65,0.87, 0.60 for GAM, RT, RF, and ANN, respectively. The RF model shows the largest improvement (47%) on R^2over the benchmark quadratic equation. The quadratic equation also has a poor MAE value (0.59 mg/l). The MAE values for these four ML models are 0.48, 0.45, 0.27, and 0.51 mg/l, respectively. The RF model also has the largest improvement (54%) over the quadratic equation.

Similarly, the RF model also outperforms other models in terms of correlation, RMSD, and SD (Fig. 6). It has a notably higher correlation coefficient, a lower RMSD, and a more desirable SD. The performance of the remaining ML models is similar to each other but consistently better than that of the benchmark model, even though the

Fig. 5 Taylor Diagrams showing the correlation coefficient (the azimuth position), standard deviation (radial distance from the origin), and centered root mean square difference (RMSD; radial distance from the reference point A which is the origin for RMSD) between the observed nitrate concentration and simulated nitrate concentration via five models consisting of the Quadratic Equation, GAM, RT, RF, and ANN. Reference point A shows the statistical metrics of the observations (RMSD = 0; R = 1)

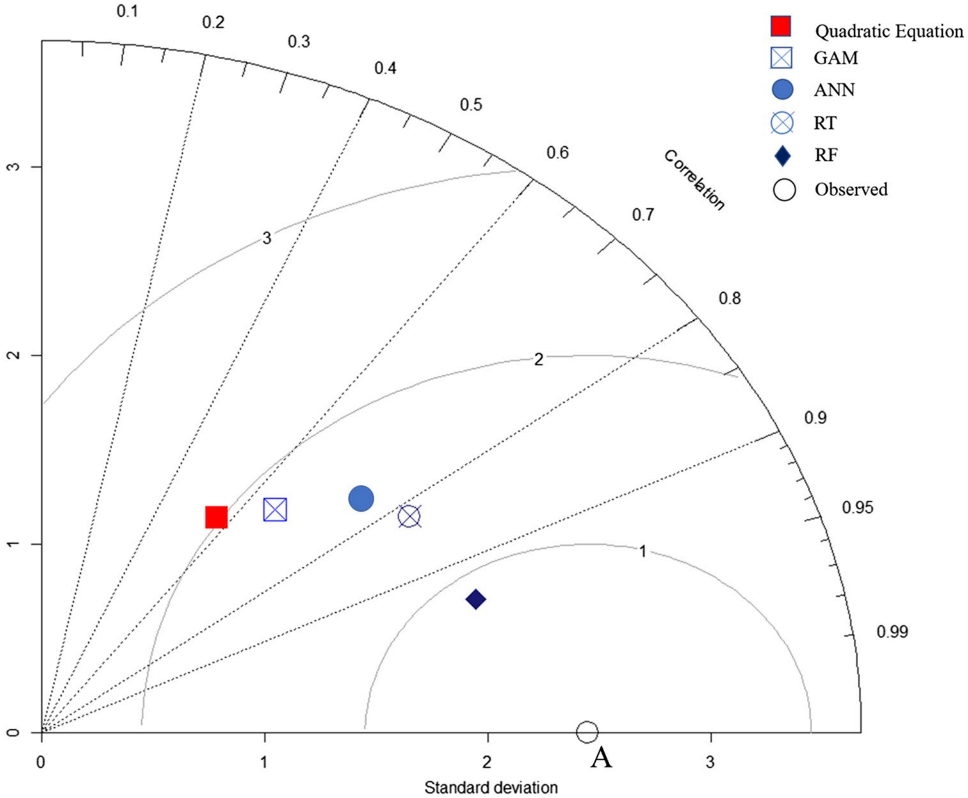
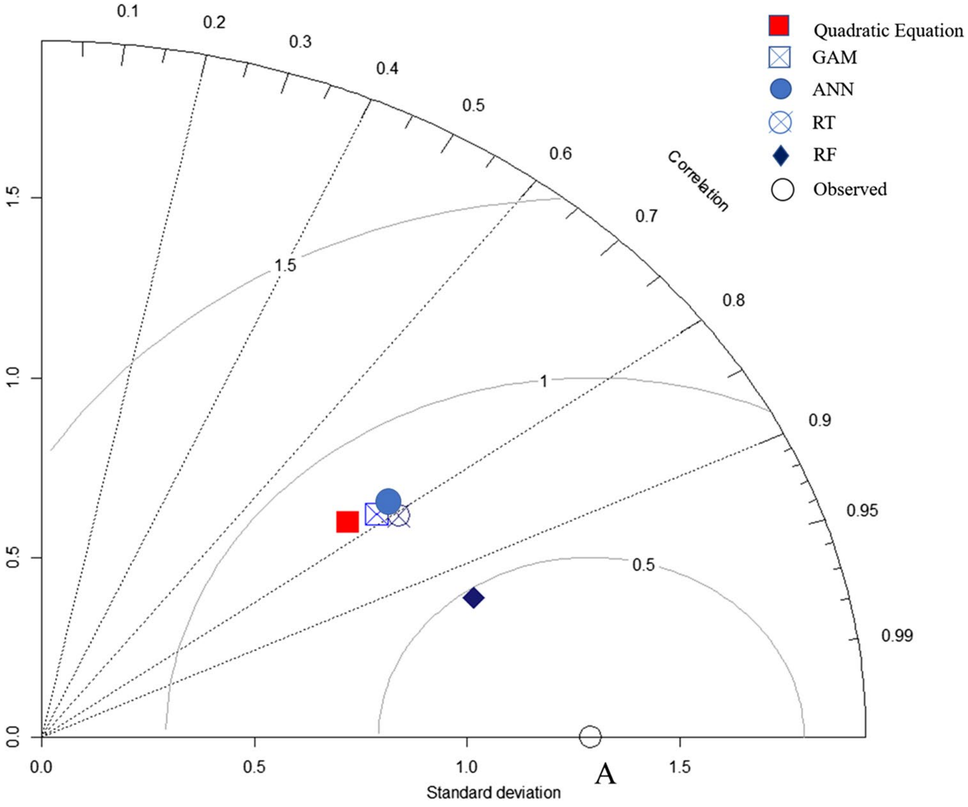


Fig. 6 Taylor Diagrams showing the correlation coefficient, standard deviation, and centered root mean square difference between the observed potassium concentration and simulated potassium concentration via five models consisting of the Quadratic Equation, GAM, RT, RF, and ANN



improvement over the benchmark model is not remarkable. All in all, based on all metrics examined, the proposed ML models are able to provide improved simulations of potassium concentration over the benchmark model. The improvement is the most noticeable in the RF model.

Simulation of boron (B)

Boron is the third ion constituent that the quadratic equation could not simulate with desirable accuracy. The R^2 values are 0.74, 0.84, 0.89,0.96, and 0.87 between the observations and simulations from the quadratic equation, GAM, RT, RF, and ANN, respectively. The corresponding MAE values are 0.06, 0.05, 0.03, 0.02, and 0.04 mg/l for these models, respectively. The RF model yields the highest R^2 and lowest MAE. Other ML models also outperform the benchmark model.

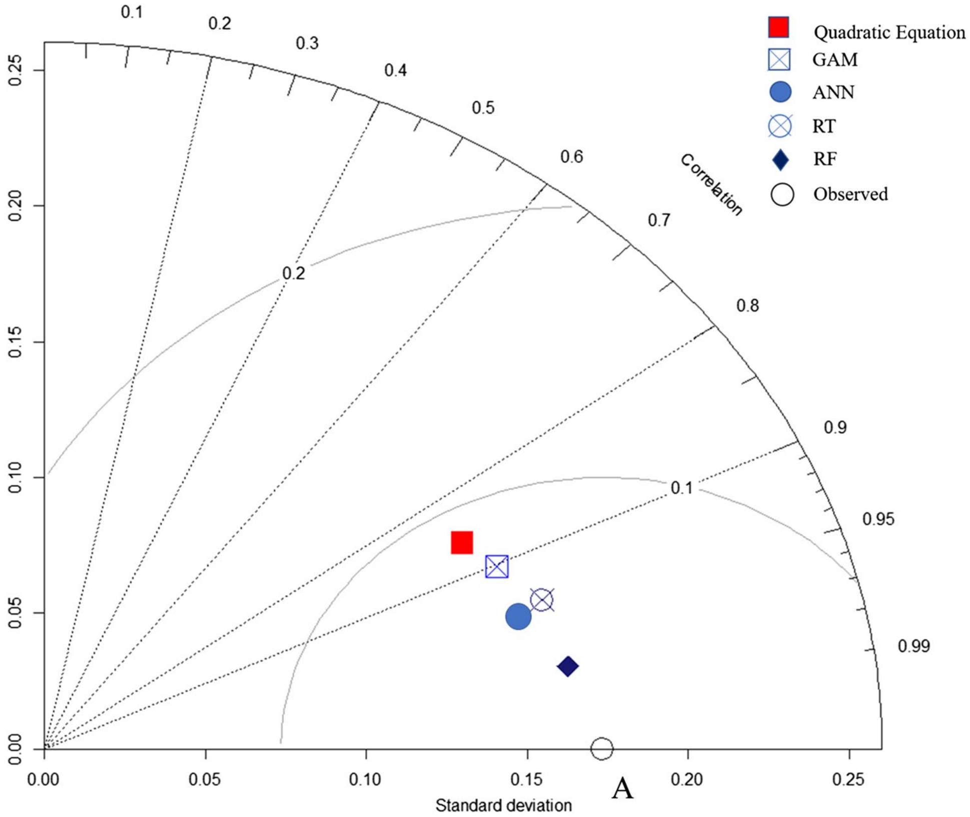
Based on the Taylor diagram (Fig. 7), the RF model also has the highest correlation, lowest RMSD, and most desirable SD among all five models. ANN and RT provide fairly close performance. The GAM model has the least desirable performance among ML models, but it still outperforms the benchmark model.

Model selection and testing the second scenario

Based on the results discussed so far, ML models consistently outperforms the quadratic equations in simulating nitrate, potassium, and boron. Among four ML models examined, the RF model has the most desirable performance. Hence, the RF model is further tested with four predictors (second scenario) by adding month and water year type as additional input features to the model.

There are 12 months (January, February, …, December) and three water year types (below normal for 2018, wet for 2019, and dry for 2020). Figure 8 displays a comparison between two scenarios (RF_2 and RF_4) based on two performance criteria that are R^2 (Fig. 8a) and MAE (Fig. 8b), respectively. Based on the results, R^2 between simulated and observed nitrate are 0.88 and 0.95 (8% improvement) for RF_2 and RF_4, respectively. Also, RF_4 decreases the MAE of simulated nitrate from 0.66 to 0.33 mg/l (50% improvement over RF_2). The RF_4 also shows improvement when applied to potassium simulation. The RF_4 increases R^2 by 2.3% and decreases MAE by 22% when compared with RF_2 results. Moreover, for boron, the RF_4 increases the R^2 by 1% and decreases MEA by 15% when compared to RF_2. Overall, these observations indicate that the second scenario (RF_4) yields better model performance.

Fig. 7 Taylor Diagrams showing the correlation coefficient, standard deviation, and centered root mean square difference between the observed boron concentration and simulated boron concentration via five models consisting of the Quadratic Equation, GAM, RT, RF, and ANN



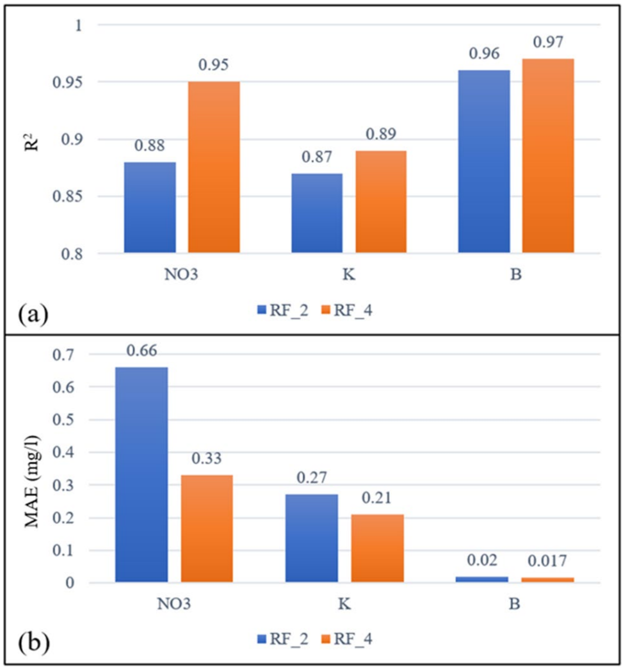


Fig. 8 RF model performance on simulating the concentrations of nitrate, potassium, and boron under two scenarios (scenario 1, RF_2, with two predictors: EC and station; scenario 2, RF_4, with four predictors: EC, station, month, and water year type) based (**a**) R^2 and (**b**) MAE

Model assessment

Out of the four proposed ML models, the RF model tends to yield the most desirable metrics. This sub-section aims to assess the prediction error (i.e., generalization error) of the RF model on new data. The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is extremely important in practice since it guides the choice of learning method or model and gives us a measure of the quality of the ultimately chosen model. Test error, also referred to as generalization error, is the prediction error over an independent test sample. The best approach for the problem is to randomly divide the dataset into two parts: a training set and a test set. The training set is used to fit the models; the test set is used to assess the generalization error of the final chosen model.

It is difficult to give a general rule on how to choose the number of observations in the training and test set, as this depends on the signal-to-noise ratio in the data and the training sample size. In this study, the data split is 80% (146 samples) for training and 20% (37 samples) for testing. In this case, the RF-4 models are tested with the independent datasets not used in the model training process to avoid overtraining (or overfitting) the models. Figure 9 shows how well the RF model performs with training and independent test datasets by comparing observed NO3, K, and B levels and their counterparts simulated via the RF models. The x-axis shows the observed data, and the y-axis shows the simulated data. The dashed line in the graphs is the 1:1 line that shows a perfect model that can simulate the observations without any errors. The quantitative performance of the RF models with the training and the independent testing datasets is also summarized in Table 4. The R^2 are 0.95 and 0.95 for training and independent testing for NO3, respectively. The MAE is 0.39 and 0.48 mg/l for training and independent testing for NO3, respectively. The simulation model for potassium has close R^2 and MAE for training and independent testing. R^2 values are 0.88 and 0.86, and MAE values are 0.24 and 0.25 mg/l for training and independent testing, respectively. The performance of the boron simulation model for training and independent testing is satisfactory. Specifically, R^2values are 0.97 and 0.97, and MAE values are 0.02 and 0.02 mg/l for training and independent testing, respectively. Overall, the results indicate that the model performance during the training process is fairly similar to its counterpart during the independent testing process. The possibility of model overfitting is remote.

Table 4 shows the performance of the RF-4 model for all 12 ion constituents based on five criteria (R^2, MAE, RMSD, Bias, and SD) for training and independent testing. Overall, R^2 for training and independent dataset for other nine ion constituents are close to 1. Also, the MAE and RMSD of the ion constituents are noticeably decreased when compared with benchmark models.

Testing the selected model on other ion constituents

The performance of the RF models in both RF_2 and RF_4 scenarios are further assessed for other nine ion constituents. For illustration purpose, Fig. 10 presents the percent improvement of the RF_2 and RF_4 models when compared with the benchmark model (quadratic equation) based on R^2 and MAE, respectively. The improvement in R^2 is between 0.2 and 3.2% for nine ion constituents. The RF models do not significantly improve the accuracy based on R^2 because the benchmark models already yield satisfactory R^2 for these nine constituents (Table 3). In contrast, the improvement in MAE is remarkable. For instance, RF_4 increases R^2 by 0.2%, but it reduces MAE by 75% over the quadratic equation for simulating TDS. Moreover, RF_4 improves MAE by 60%, 66%, 59%, 60%, 46%, 60%, 53%, and 48% for Cl^-, $SO4^{2-}$, Na^+, Ca^{2+}, Mg^{2+}, Br^-, hardness, and Alkalinity, respectively. These observations indicate that though the quadratic questions can yield fairly reasonable simulations on these nine constituents, the RF models (particularly with four predictors) can yield even better simulations with notably smaller errors (measured by MAE).

Fig. 9 Observed (x-axis) RF model-simulated (y-axis) on the concentrations of nitrate (first row), potassium (second row), and boron (third row). The first column (panels (**a**), (**b**) and (**c**)) and second column (panels (**d**), (**e**), and (**f**)) show the training and validation results, respectively

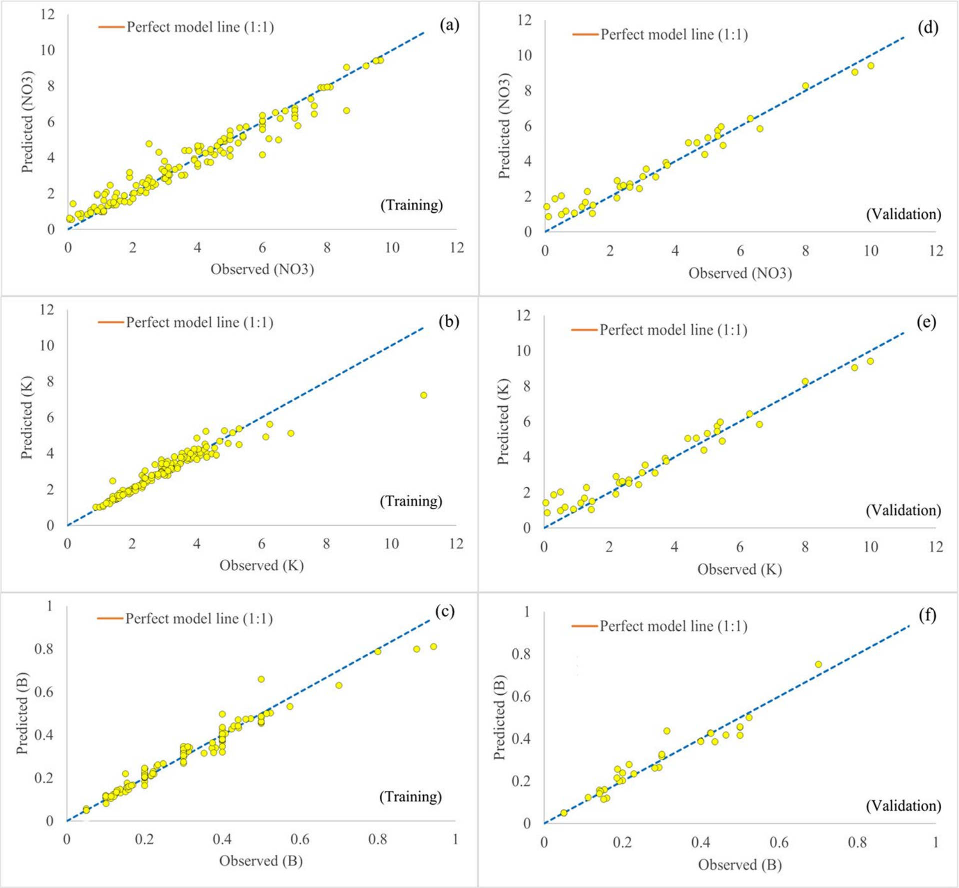


Table 4 Performance of the RF model with four predictors

Ion Constituents	R^2		MAE		RMSD		Bias		SD	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
B	0.97	0.97	0.02	0.02	0.03	0.03	-0.1	-2.5	0.16	0.15
Br^-	0.99	0.99	0.02	0.02	0.02	0.03	-0.2	-0.3	0.33	0.18
Ca^{2+}	0.997	0.998	1.33	1.29	1.98	1.74	0	2.3	34.67	33.34
Cl^-	0.998	0.998	3.28	3.14	5.53	4.47	-0.2	1.9	107.92	89.51
Hardness	1.00	1.00	4.79	5.06	6.96	6.71	-0.1	0.8	144.07	134.94
K	0.88	0.86	0.24	0.25	0.47	0.44	0.2	1.4	1.14	1.13
Mg^{2+}	1.00	1.00	0.49	0.61	0.74	0.90	0	1.8	14.28	13.93
NO3	0.95	0.95	0.39	0.48	0.55	0.66	0.7	3	2.13	2.54
Na^+	1.00	1.00	1.98	2.54	3.19	3.13	-0.2	-0.6	51.12	43.49
$SO4^{2-}$	1.00	1.00	2.38	2.36	3.96	3.19	-0.1	0	58.28	53.04
TDS	1.00	1.00	7.14	8.25	11.94	12.41	0.2	0.6	314.82	298.75
Alkalinity	0.99	0.98	2.81	3.94	4.20	5.71	0	-1	43.57	40.83

Discussions and conclusions

Findings

This study developed four types of Machine Learning (ML) models (i.e., GAM, RT, RF, and ANNs) within the R statistical computing environment to simulate the concentrations of 12 ion constituents in the South Delta. The results are compared to those of the conventional quadratic regression equations previously developed. The key findings are summarized as follows:

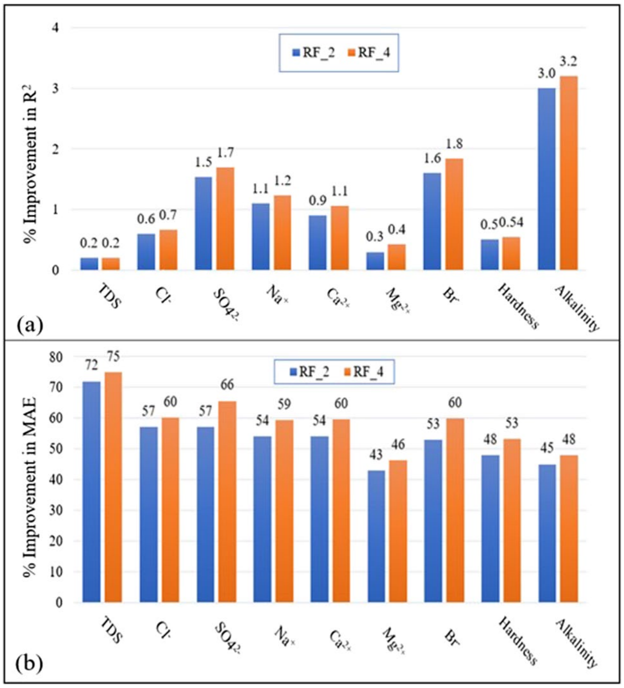


Fig. 10 RF model performance on simulating the concentrations of nine ion constituents under two scenarios (RF_2 and RF_4) based on percent improvement (from the benchmark model) in (**a**) in R^2 and (**b**) MAE

- ML models showed comparable or better performance in simulating the concentrations of ion constituents than the conventional quadratic equations.
- Among all ML models, the RF models tended to yield the best performance metrics.
- Using additional input features such as station name as well as and the corresponding time (month and the type of the year when the samples were collected) as categorical variables improved the performance of the RF models.
- RF models by design minimize the potential of model overfitting, which was confirmed in this study by testing the trained models using randomly selected independent datasets.

Implications, limitations, and future work

This study has both important scientific and practical implications. From a scientific standpoint, the study was the first to develop and apply machine learning models to simulate ion constituents given salinity in the Sacramento-San Joaquin Delta region. Quadratic regression equations have been traditionally utilized for that purpose. However, given the physical complexity of the Delta along with stringent flow and salinity regulations in the Delta, the relationships between some constituents and salinity might not be adequately represented via quadratic equations. This study exemplified the feasibility and applicability of ML models in better capturing these relationships than the traditional quadratic equations. This laid foundation for further exploration of more state-of-the-art machine learning techniques (e.g., deep learning) in this region.

From a practical point of view, the findings of the study can inform future sampling practices in the South Delta. For instance, even though the ML models generally outperform the quadratic equations in simulating all 12 ion constituents, there are a few constituents (e.g., Potassium, Nitrate, and Boron) on which the performance of the ML models can be further improved. One way is to collect additional samples for these constituents at more locations. Another way is to also collect relevant ancillary information that affects the nutrient update as well as the mix and dispersion of these constituents. These factors may include surface and subsurface aquatic vegetation coverage, flow direction and flow rate at the sampling locations, water diversions, among others. In addition, compared to the traditional quadratic equations, the ML models can yield better estimations of the concentrations of these ion constituents at times when the samples are not available and thus better inform water planning and management operations.

Despite its scientific and practical values, this study was meant to be a proof-of-concept study using the most recent grab sample dataset which has a limited sample size. Taking grab samples and processing them are labor intensive. California Department of Water Resources started collecting grab samples in the South Delta decades ago. However, earlier samples were generally sparse temporally and spatially. Moreover, not all 12 ion constituents were analyzed from these samples. In addition, sample collection and analysis methods have also evolved. Furthermore, there have been changes (e.g., in land use and local drainage, land subsidence, island flooding, seawater intrusion, etc.) in the Delta over the years. Collectively, these factors affect the accuracy and representativeness of the samples collected. This study focused on the most recent sample dataset that most accurately reflects the salinity-constituent relationships in the current Delta. The sample collection method was consistent, so were the analysis method for all 12 constituents. In a follow-up study, all available historical grab sample data across the Delta are being assembled. The ML models developed in the current study will be applied to that dataset. The resulting outcomes will be compared to that of the current study to assess the sensitivity of the ML models to data from different periods collected via likely different collection and/or analysis methods.

It is also worth noting that the ML models developed in the current study are all single task learning based (i.e., one model per ion constituent). Given that some of these ion constituents are intercorrelated with one another, as illustrated in Fig. 3. It may be beneficial in terms of efficiency and effectiveness to train ML models for all or subgroups constituents altogether. In our future work, we will explore the multi-task learning based ML approach to simulate all or subsets of 12 ion constituents simultaneously via a single model.

Concluding remarks

This study developed machine learning models to emulate the conventional regression equations in simulating ion constituents in south Sacramento-San Joaquin Delta of California. The study indicated that the Random Forest models have the most desirable performance among four machine learning models proposed. The study further discussed the scientific and practical values along with the limitations of these models. The study also described future work to further improve these models and apply them to other areas in the Delta. In a nutshell, this study suggested that machine learning can supplement the regression equations to inform water resources planning and management practices in terms of providing more robust estimations on constituents given salinity, especially for constituents of which the relationships with salinity could not be adequately represented by regression equations.

Acknowledgements We gratefully thank our colleagues in the North Central Region Office for collecting and analyzing the grab samples. The views expressed in this paper are those of the authors, and not of the State of California.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Peyman Namadi and Minxue He. The first draft of the manuscript was written by Peyman Namadi, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

The authors have no relevant financial or non-financial interests to disclose. The dataset used in this study is available at: https://wdlbeta.water.ca.gov/WaterQualityDataLib.aspx. The source code of the machine learning models developed is available at the following GitHub link: https://github.com/PeymanHNamadi/South-Delta-Ion.

References

Becker L, Sparks D, Fults D, Yeh W (1976) Operations models for central valley project. J Water Resour Plan Manag Div 102(1):101–115

Bennett W (2005) Critical assessment of the delta smelt population in the San Francisco Estuary, California. San Francisco Estuary Watershed Sci 3(2)

Breiman L, Friedman J, Stone C, Olshe R (1984) Classification and regression trees. Chapman & Hall/CRC, Boca Raton, FL, USA

Breiman L (1999) Random forests. UC Berkeley TR567

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

California Urban Water Agencies (1995) Study of Drinking Water Quality in Delta Tributaries. Report prepared by Brown and Caldwell, Archibald & Wallberg Consultants, Marvin Jung & Associates, and McGuire Environmental Consultants, Inc

California Department of Water Resources (CDWR) (2012) Estimating delta-wide bromide using DSM2- Simulated EC Fingerprints. 33rd Annual Progress Report. CDWR, Sacramento

Candel A, Viraj P, Erin L, Anisha A (2016) Deep learning with H2O, H2O. ai Inc

Chen L, Roy S (2015) Generalized delta conservative constituent modeling using artificial neural networks

CSWRCB (1999) Water right Decision 1641. CSWRCB, Sacramento, 225

Denton R (2015) Delta salinity constituent analysis. Richard Denton and Associates, prepared for the State Water Project Contractors Authority

Guivetchi K (1986) California Department of Water Resources Interoffice Memo, Salinity unit conversion equations https://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/california_waterfix/exhibits/docs/petitioners_exhibit/dwr/dwr_316.pdf

Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 3:297–318

Hastie T, Tibshirani R, Friedman J, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, New York, pp 1–758

He M, Zhong L, Sandhu P, Zhou Y (2020) Emulation of a process-based salinity generator for the Sacramento–San Joaquin Delta of California via deep learning. Water 12(8):2088

He M, Anderson J, Lynn E, Arnold W (2021) Projected changes in water year types and hydrological drought in California's Central Valley in the 21st Century. Climate 9(2):26

Healey M, Dettinger M, Norgaard R (2016) Perspectives on bay–delta science and policy. San Franc. Estuary Watershed Sci 2016:14

Ho T (1995) Random decision forests. In: Proceedings of the 3rd international conference on document analysis and recognition. Montreal, QC, 14–16 August 1995, pp 278–282

Hopfield J (1988) Artificial neural networks. IEEE Circ Devices Mag 4(5):3–10

Hutton P (2006) Validation of DSM2 volumetric fingerprints using grab sample mineral data. Power Point presentation at CWEMF Annual Meeting, March 2006

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. springer, New York, p 18

Kumar M, Puri A (2012) A review of permissible limits of drinking water. Indian J Occup Environ Med 16(1):40

Lewis W, Donald M (1986) Toxicity of nitrite to fish: a review. Trans Am Fish Soc 115(2):183–195

Li P, Wu J (2019) Drinking water quality and public health. Expos Health 11(2):73–79

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:3

Loh W (2011) Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov 1(1):14–23

Lopez P, Pérez-Rodríguez I, Estrany F, Devesa R (2017) Effects of sulfate and nitrate on the taste of water: a study with a trained panel. J Water Supply: Res Technol —AQUA 66(8):598–605

McCulloch S, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115–133

Mohsin M, Safdar S, Asghar F, Jamal F (2013) Assessment of drinking water quality and its impact on residents' health in Bahawalpur city. Int J Humanit Social Sci 3(15):114–128

Moyle P, Hobbs A, Durand J (2018) Delta Smelt and water politics in California. Fisheries 43(1):42–50

Namadi P, Deng Z (2021) Modeling and forecasting vibrio parahaemolyticus concentrations in oysters. Water Res 189:116638

NCRO South Delta Ion Report. Technical Report (2021) California Department of Water Resource North Central Region Office, Sacramento

Potdar K, Taher P, Chinmay P (2017) A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications.

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Sabet M, Coe J (1986) Models for water and power scheduling for the California State water project 1. JAWRA J Amer Water Resour Assoc 22(4):587–596

Suits B (2002) Chap. 5, Relationships between Delta Water Quality Constituents as derived from Grab Samples. In: DWR's "Methodology for Flow and Salinity Estimates in the Sacramento-San Joaquin Delta and Suisun Marsh." 23rd Annual Progress Report, June 2002

Taylor K (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Research: Atmos 106(D7):7183–7192

Therneau T, Atkinson B (2019) rpart: Recursive Partitioning and Regression Trees. R package version 4.1–15. 2019

USFWS (2008) Formal endangered species act consultation on the proposed coordinated operations of the Central Valley Project (CVP) and State Water Project (SWP). USFWS, Sacramento, p 410

World Health Organization (2008) Guidelines for drinking-water quality. Third edition Incorporating the first and second addenda, p 1

World Health Organization (2009) Boron in drinking-water: Background document for development of WHO Guidelines for Drinking-water Quality. No. WHO/HSE/WSH/09.01/2. World Health Organization

World Health Organization (2017) Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth

Wood S (2017) Generalized additive models: an introduction with R, 2nd edn. CRC

Zhang Z, Deng Z, Rusch K, Walker N (2015) Modeling system for predicting enterococci levels at Holly Beach. 109: 140 – 47. Mar Environ Res 109:140–147

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.