JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 58, No. 6 AMERICAN WATER RESOURCES ASSOCIATION December 2022

# Evaluate River Water Salinity in a Semi-Arid Agricultural Watershed by Coupling Ensemble Machine Learning Technique with SWAT Model

*Chunggil Jung, Sora Ahn, Zhuping Sheng, Essayas K. Ayana, Raghavan Srinivasan, and Dhanesh Yeganantham*

**Research Impact Statement**: Better understanding of patterns of river water salinity with machine learning leads to adaptive management and potential use of marginal quality water for agricultural production and urban supplies).

ABSTRACT: This study is to establish a new approach to estimate river salinity of semi-arid agricultural watershed and identify drivers by using hydrologic modeling and machine learning. We augmented the limitations of the Soil and Water Assessment Tool (SWAT) to model salinity by coupling with eXtreme Gradient Boosting (XGBoost), a decision-tree-based ensemble machine learning algorithm. Streamflow, precipitation, elevation, main reach length, and dominant soil texture of the top two layers were used along with $NO_3$, $NO_2$, and total phosphorus (TP) output from a calibrated SWAT model are used as predictors to Total Dissolved Solids (TDS) in the XGBoost algorithm. Then, the SWAT model simulations of streamflow, $NO_3$+$NO_2$, and TP from 2000 to 2015 are used as inputs of the XGBoost model to predict monthly water TDS distribution along the river. The predicted river water TDS showed a higher concentration as going downstream from El Paso (inlet) through the Hudspeth canal to Fort Quitman (outlet). Finally, this study carried out cause analysis focusing on soil physical characteristics. The soil salinity level is directly affected by the soil permeability and irrigation water. As a result, the highest TDS is shown in sites with silt loam, whereas the lowest TDS was shown in sites with very cobbly soil. Silt soils can hold more water and are slower to drain than soils of a sand type. These analyses can be used to better understand the mitigation of water salinity.

(KEYWORDS: watershed management; machine learning; SWAT; water salinity; soil texture; irrigation; watershed; surface water/groundwater interactions.)

## INTRODUCTION

Over the last 50 years, irrigated area had more than doubled and contributed significantly to the world agriculture output and food supply (FAO 2011). Irrigation has mainly been used to control the water content in the fields, compensate for the lack of precipitation (PCP), and suppress weed growth. Nevertheless, such expansion happened with severe consequences to the environment. The most pressing of these consequences is irrigation-induced salinity that has become an increasing problem in several countries (Umali 1993). Nearly one-third of the irrigated land worldwide is affected by salinization (Schwabe and Kan 2006) and saline environments tend to hinder agricultural production by lowering crop yields. With a predicted increased temperature, declined rainfall, and reduced snowmelt due to climate change, agriculture in the majority of arid and semi-arid regions needs to rely

Paper No. JAWR-20-0184-P of the *Journal of the American Water Resources Association* (JAWR). Received December 18, 2020; accepted October 13, 2021.  **Discussions are open until six months from issue publication.**

Texas A&M AgriLife Research at El Paso (Jung; Ahn; Sheng), El Paso, Texas, USA; and Department of Ecosystem Science and Management (Ayana, Srinivasan, Yeganantham), Texas A&M University, College Station, Texas, USA (Correspondence to Ahn: sahn@kitprofs.com).

*Citation*: Jung, C., S. Ahn, Z. Sheng, E.K. Ayana, R. Srinivasan, and D. Yeganantham. 2022. "Evaluate River Water Salinity in a Semi-Arid Agricultural Watershed by Coupling Ensemble Machine Learning Technique with SWAT Model." *Journal of the American Water Resources Association* 58 (6): 1175–1188. https://doi.org/10.1111/1752-1688.12958.

even more on irrigation (Postel 1999; Ragab and Prudhomme 2002; Adhikari and Nejadhashemi 2015).

Salinity is a significant factor limiting the usability of the water in El Paso of the Lower Rio Grande Basin (Moyer et al. 2009). Salinity increases dramatically from 40 mg/L in the headwater in Colorado to 2,000 mg/L in a series of stretches along the river in the Lower Rio Grande Basin (Phillips et al. 2003; Hogan et al. 2007). The influences of natural sources are more evident at lower flows. The operational schedule of the river results in higher salinity concentrations during times of reduced releases of freshwater in the winter nonirrigation season (Doremus and Lewis 2008; USACE 2011). Like the surface water quality, groundwater quality in El Paso has also deteriorated with depth and to the southeast (Sheng 2013). The groundwater level declines caused by pumping have resulted in brackish groundwater intrusion in the El Paso area, and the loss of several municipal supply wells (Sheng and Devere 2005; Montegomery & Associates and Hutchison 2016). This deterioration of groundwater quality is mainly characterized by increased chloride and Total Dissolved Solids (TDS) concentrations (Heywood and Yager 2003; Hutchison 2004). To better understand the pattern and trend of the salinity condition we need to improve our capacity for simulating the salinity in the river. In turn, it will help us develop guidelines for salinity management and better uses of marginal quality water, securing water supplies for future agricultural production and municipal water supplies across the United States (U.S.) and Mexico Border.

Interests in surface and groundwater salinity modeling using a variety of approaches (numerical modeling, stochastic analysis, and machine learning) have steadily grown. For example, the Hydrus and MODFLOW numerical models coupled with MT3D have been used to evaluate salinity dynamics (Hanson and Hopmans 2008; El-Bihery 2009; Kanzari et al. 2012; Ibrahimi et al. 2014; Eissa et al. 2016; Ghorbani et al. 2017). More recently, Wu et al. (2018) presented spatial distribution and severity of salinity by combining satellite and radar datasets using machine learning. Moreover, Vermeulen and Niekerk (2017) reported the most accurate algorithm for the prediction of salinity among various machine learning algorithms using ensemble machine learning algorithms.

Various numerical models have been developed to simulate the more complex process in the system of water, soil, crop, and salinity. But the models require large numbers of input parameters as big data. Such model simulations often require considerable time and effort to compile input data and larger computation resources that can handle the increasing refinement and complexity of numerical models (Chen et al. 2020). In contrast, machine learning approaches have been applied over the past decades for simulating various hydrological processes including water dynamics and water quality with significant prediction accuracy (Karandish and Simunek 2016). The prediction capability of the machine learning algorithms is limited by the information contained in the data and they do not enable intuitive interpretation (Lamorski et al. 2013) of the evaluated processes. Nevertheless, machine learning is the most popular algorithm nowadays that can overcome the disadvantages of other modeling approaches. Coupling machine learning and numerical modeling in such a way that one complements the weakness of the other could offer a significant improvement in prediction accuracy while also maintain a reasonable representation of complex processes (Vandenberghe et al. 2007; Tuv et al. 2009). The main purpose of this study was thus to predict spatiotemporal river water TDS as the indicator of salinity in the semi-arid agricultural watershed of the Lower Rio Grande using a machine learning algorithm with monitoring data and Soil and Water Assessment Tool (SWAT) model output. The specific objectives of the study are as follows: (1) to develop an eXtreme Gradient Boosting (XGBoost) algorithm to estimate the river TDS using observed salinity monitoring data; (2) to evaluate the accuracy of developed the XGBoost algorithm after optimization of parameters; (3) to simulate streamflow and water quality components with the calibration process of SWAT model in the whole river by replacing observed data with SWAT-calibrated results; and (4) to distribute river salinity from SWAT output results as input variables for machine learning algorithm and identify the reasons why the river TDS changes along the river.

## MATERIAL AND METHODS

### *Characteristics of Study Watershed*

In this study, the Rio Grande watershed from El Paso Gaging Station to Fort Quitman Gaging station was selected as the study site in the Far West Texas (Figure 1), which is located within the latitude range of 30°48′9″ N to 31°59′37″ N and longitude range of 106°39′12″ W to 107°24′55″ W. The study watershed includes the Hueco Bolson aquifer. The Hueco Bolson aquifer spans about 2,500 square miles (6,475 $km^2$), or 1.6 million acres in New Mexico, Texas, and Chihuahua (Figure 1). The study area includes all of the El Paso County, Texas, and adjacent portions of Doña Ana and Otero counties, New Mexico, and Hudspeth County, Texas. The metropolitan areas of El Paso, Texas, and Ciudad Juarez, Chihuahua, Mexico

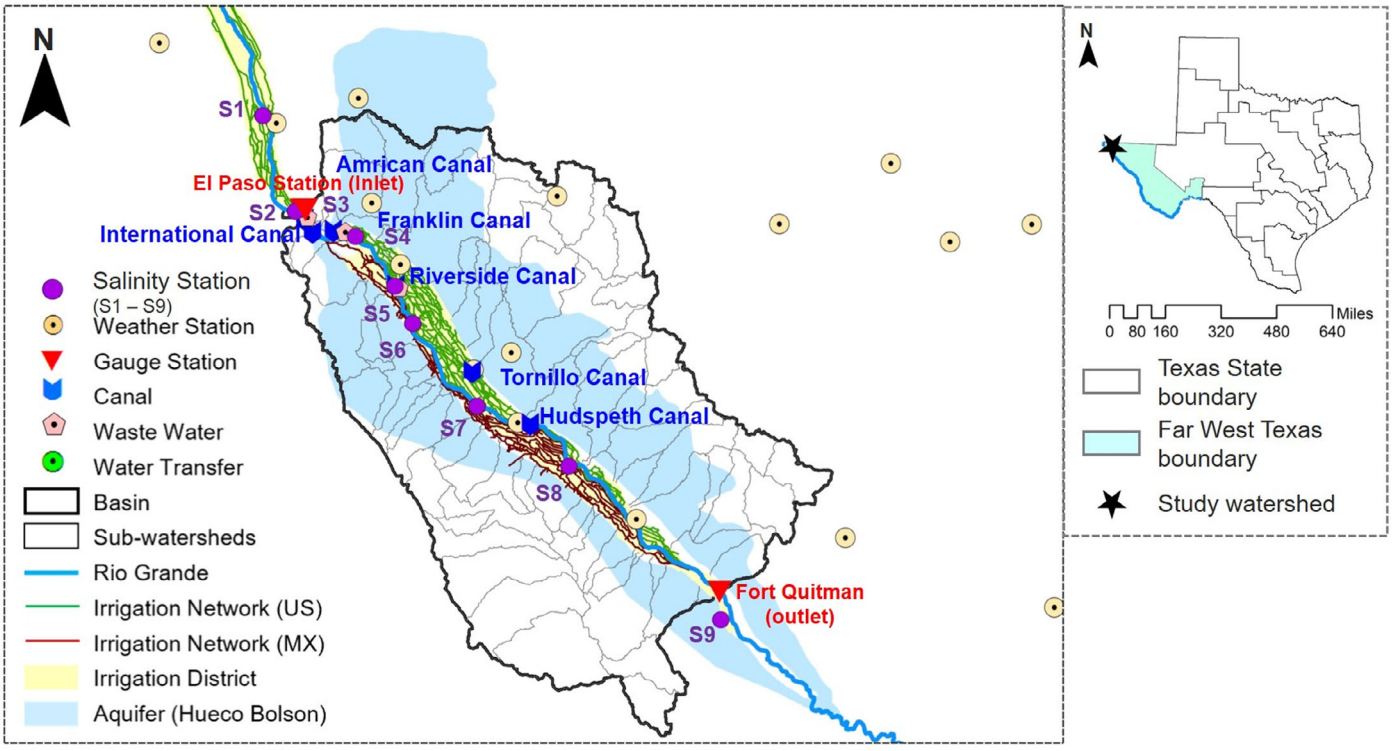


FIGURE 1. Location of the study watershed and aquifer including irrigation networks, canals, and gauging stations for the weather, streamflow, and water quality. MX, Mexico; US, United States

located in the study area (Montegomery & Associates and Hutchison 2016).

### *Observed Data for Development of XGBoost Algorithm and SWAT Model*

In the study area, there are fourteen weather stations of the National Oceanic and Atmospheric Administration (NOAA), two streamflow gauging stations at the watershed inlet (El Paso) and outlet (Fort Quitman), and nine water quality gauging stations from S1 to S9 along the river channel operated by International Boundary and Water Commission (IBWC). Within the study watershed, six major canals (American, International, Franklin, Riverside, Tornillo, and Hudspeth) located along the river provide irrigation water in the crop areas of the U.S. and Mexico (Figure 1). The American Diversion Dam diverts water into the International Diversion Dam, about 3.2 km below the American Diversion Dam, for the Mexican side of the El Paso Valley. The 3.2-km-long American Canal diverts the water into the Franklin Canal and Riverside Canal to use for irrigation of the 140-km-long El Paso valley on the U.S. side.

SWAT model requires elevation, land use, soil, and meteorological data corresponded to the study area. Figure 2 shows the elevation, land use, and soil information for the study area. The study area is a semi-arid climate. The average annual PCP of the study area is about 250 mm, most of which occurs in the summer months (Heywood and Yager 2003). The elevation ranged from 1,034 to 2,115 m, and the average is 1,288 m (Figure 2a). The watershed area is about 7,988 km$^2$. As for land use, the crops which are classified into 10 types have been cultivated along Rio Grande (Figure 2b). Pecans and cotton are dominant in the crops. The predominant soil types in the watershed are very gravelly (26.7%), very fine sandy loam (22.9%), and loamy fine sand (19.8%) (Figure 2c). And the soil texture of the area usually has high permeability (Montegomery & Associates and Hutchison 2016).

The monthly water quality data for 18 years (1998–2015) were obtained from IBWC at nine stations (S1–S9). The average monthly values of TDS represent two distinct seasons in a year that significantly differ from each other. The season from October to February represents a high TDS concentration period with an average TDS of 1,463 mg/L, while the season from March to September has lower concentrations of TDS with an average value of 763 mg/L. The difference between them can be explained by the variation of river flow. Flow with high volumes helps decrease TDS concentration and accordingly TDS concentration tends to be high during a nonirrigation season when the river flow is relatively low (Abudu

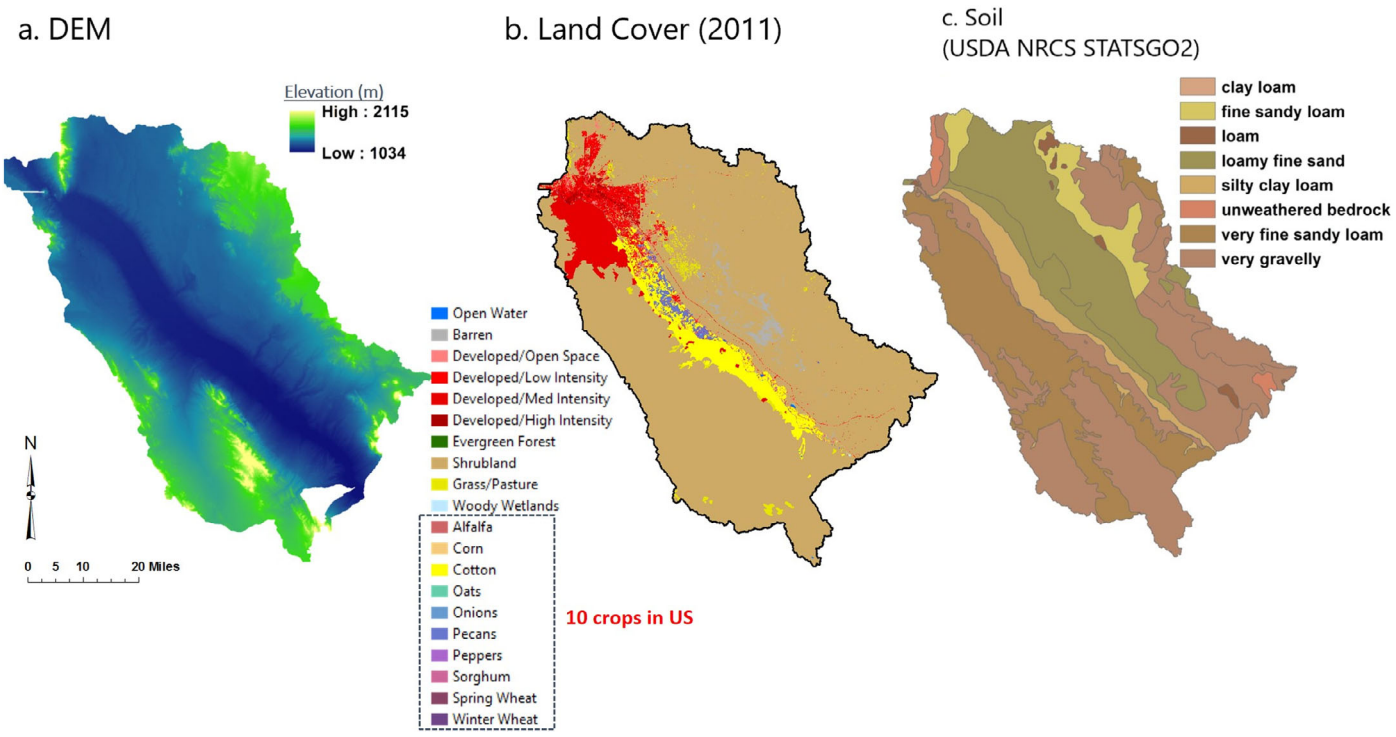


FIGURE 2. SWAT topographical input data: (a) elevation (DEM), (b) land cover map including 10 crops, (c) soil texture map for the study watershed. DEM, Digital Elevation Model; SWAT, Soil and Water Assessment Tool.

and King 2012). From the observed stations, the average chloride is 227.4 mg/L. The others are 7.6 $m^3/s$ for flow, 1.7 mg/L for $NO_3$+$NO_2$, 0.8 mg/L for total phosphorus (TP), 173.4 mg/L for total suspended solids (TSS). The flow and water quality data used for specifying independent variables through the correlation analysis include streamflow, TDS, chloride, $NO_3$+$NO_2$, and TP.

### *Model Configuration*

The correlation analysis has been conducted using Pearson's coefficient to identify which variables in the observed data are contributing to the salinity (Table 1). Correlations among the variables were not statistically significant except chloride and TDS, because both chloride and TDS stand for the indicators which can express the sensitivity of salinity. Thus, either one of them should be used as a target variable. In the study, TDS was determined as a target variable because of good quality data and less missing data within the watershed. As seen from Table 1, no certain single variable has a high relationship with the TDS. This means that the TDS cannot be predicted using only a simple regression model with a single value. So, this study should use the more complex models in the same way that the ensemble model or a machine learning technique has a much complex approach to overcome the problem. To overcome the difficulties identified above, we develop an innovative approach for the simulation of water salinity within the study area. It couples the machine learning algorithms with the SWAT model. The following sections will explain the model configuration in more detail.

**The XGBoost Tree Algorithm.** The XGBoost is an ensemble algorithm for decision-making by mixing several tree models. The regression tree and the gradient boosting are combined into decision trees with appropriate trimming. The algorithm consists of multiple decision trees, with each tree gradient down by learning from the previous order of the tree. Finally,

TABLE 1. The Pearson's coefficient for correlation between observed water quality data.

| Correlation | Chloride (mg/L) | Flow ($m^3/s$) | $NO_3$+$NO_2$ (mg/L) | TP (mg/L) | TDS (mg/L) |
|---|---|---|---|---|---|
| Chloride (mg/L) | 1 | −0.279 | 0.070 | 0.068 | 0.852 |
| Flow ($m^3/s$) | −0.279 | 1 | −0.188 | −0.035 | −0.401 |
| $NO_3$+$NO_2$ (mg/L) | 0.070 | −0.188 | 1 | 0.192 | 0.093 |
| TP (mg/L) | 0.068 | −0.035 | 0.192 | 1 | −0.058 |
| TDS (mg/L) | 0.852 | −0.401 | 0.093 | −0.058 | 1 |

Notes: TDS, Total Dissolved Solids; TP, total phosphorus.

the aggregation of all the trees produces the final model and decision (Nalenz and Villani 2018).

**Machine Learning System — XGBoost.** Machine learning systems provide powerful algorithms. However, if the algorithm fits the data too well, the variance term is large, and hence the overall error is increased. It has been known for overfitting. XGBoost algorithm prevents the overfitting problem by conducting normalization for each model. The algorithm is also known for additive training, which learns the previous result sequentially in the current stage. In additive training previous results sequentially affect the training of the current stage (Nalenz and Villani 2018; Mehta et al. 2019). Thus, the training improves a predicted value approach to the target value iteratively. This training will learn from the weakly data, and gradually get closer to the actual value, unlike other tree models. The method has the strength in terms of the bias and the variance. Even if current data have a high bias because of weakly learning, the high bias can be sufficiently improved from weak learning as going sequentially. Besides, the variance of the predicted result also can reduce sufficiently at the same time (Srivastava et al. 2014; Mehta et al. 2019).

This method has been widely used in machine learning for hydrological and environmental problems. Cisty and Soldanova (2018) predicted time series of the river flow at where fewer input data served for the simulation by using XGBoost. White (2015) also forecasted streamflow at ungauged sites undergoing drought. As environment issues, Qin et al. (2018) studied soil hydrological status to succeed in N demand modeling for corn with the application of machine learning algorithms including XGBoost. Compared with the general linear models built with traditional approaches, most results show that the XGBoost model achieves better time-series results in the field of water resource and environment.

**Algorithm Structure in This Study.** The XGBoost algorithm in this study is used for its sequential modeling with weakly learning data compared to simple random forest algorithms. This algorithm has the advantage of lowering the bias by mixing underfitting models (low variance) and learning more trees sequentially. This can improve the variance of predicted results due to the overfitting of training data, which is known for the biggest vulnerability to machine learning. Also, in many existing studies, XGBoost proved to perform better than Random Forest when optimal parameter tuning is performed (Punnoose and Ajit 2016; Zhang et al. 2018). To optimize the XGBoost, the structure of the XGBoost is as follows: (1) data splitting (training and test), (2) $k$-fold cross-validation, (3) optimal parameter tuning using grid-search, and (4) prediction of the results.

In machine learning, two tasks are commonly done at the same time in data pipelines: cross-validation and (hyper) parameter tuning. Cross-validation is the process of training learners using one set of data and testing it using a different set. Parameter tuning is the process to select the values for a model's parameters that maximize the accuracy of the model. As a cross-validation, $k$-fold cross-validation was used as a training method. The $k$-fold cross-validation separates the training data into "$k$" folds without overlap. One set of $k$ folds is separated by training and validation. Hence, $k$ models are made, and the models are trained until $k$ times. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the results are averaged over the rounds to estimate a final predictive model. The next step for optimizing parameters, Grid Search library exhaustively considers all parameter combinations for optimal parameters. The library implements a "fit" method and a "predict" method like any classification or regression except that the parameters used to predict are optimized by cross-validation. The Grid Search library consists of an estimator (regressor or classifier), a parameter space, a method for searching or sampling candidates, a cross-validation scheme, and a score function. During the call for the Grid Search library to fit, it selects the parameters on a specified parameter grid by maximizing a score (the scoring method of the underlying estimator). The prediction, score, or transform is then delegated to the tuned estimator.

**Description of SWAT.** SWAT is a physically based, continuous, long-term, distributed-parameter model designed to predict the effects of land management practices on hydrology and water quality in agricultural watersheds under varying soil, land use, and management conditions (Arnold et al. 1998). SWAT is based on the concept of hydrologic response units (HRUs), which are portions of a subbasin with unique land use, management, and soil attributes. The runoff, sediment, and nutrient loadings from each HRU are calculated separately based on weather, soil properties, topography, vegetation, and land management and are then summed to determine the total loading from the subbasin (Neitsch et al. 2009; Park et al. 2011, 2014).

**Linkage of XGBoost Algorithm with SWAT Model.** At the last step, we coupled the machine learning algorithms, namely XGBoost with the SWAT model. This approach was used to predict river TDS by using a calibrated SWAT model outputs as an input to a machine learning algorithm (Figure 3). Among these processes, the SWAT model simulated

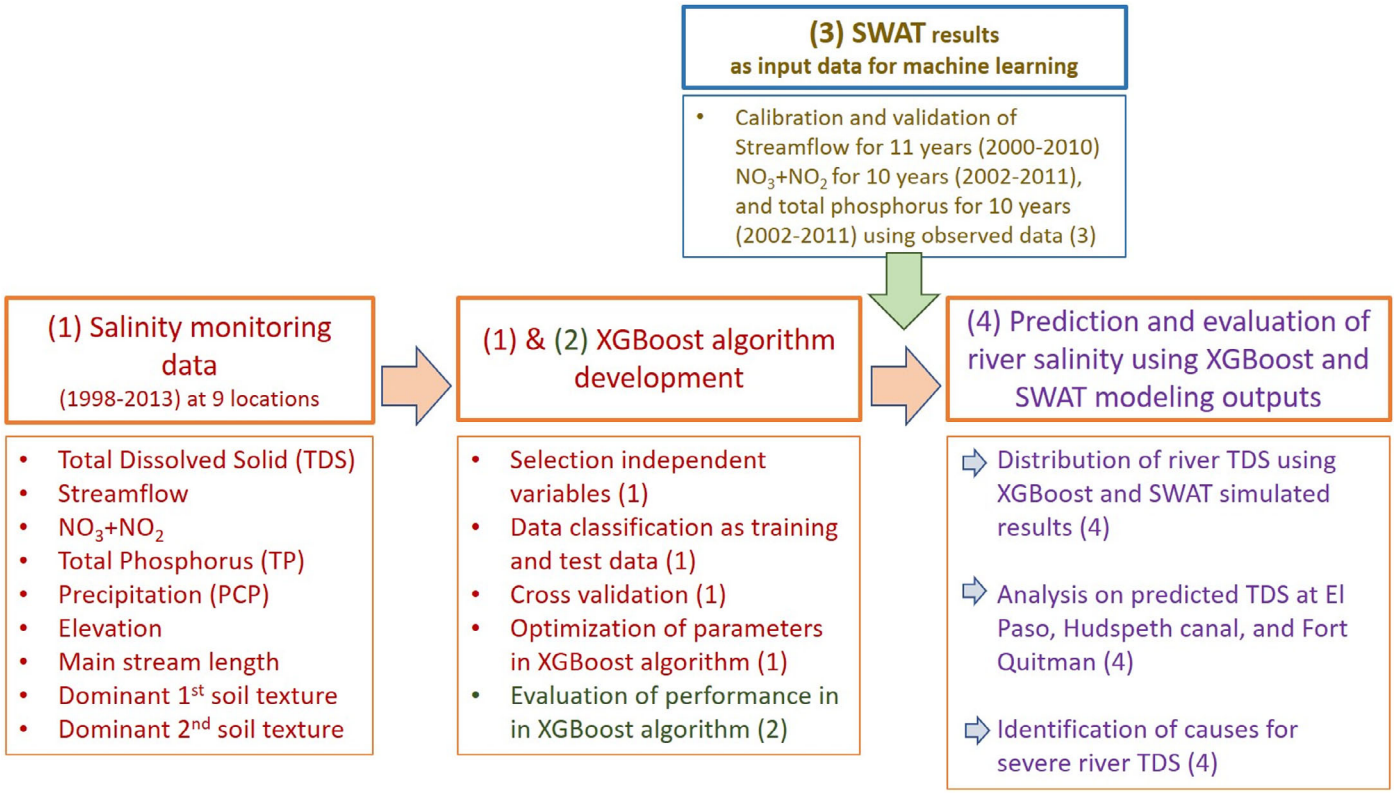


FIGURE 3. Flowchart of the study in terms of linkage between the eXtreme Gradient Boosting (XGBoost) algorithm and the SWAT model for implementation of the XGBoost algorithm and SWAT model.

both all hydrological and water quality variables, which can be spatially distributed, however, the current SWAT model cannot directly simulate TDS due to the unavailability of the required data. Therefore, a machine learning approach that can predict TDS from the flow and water quality is implemented by employing distributed SWAT output in lieu of observed independent variables.

## RESULTS AND DISCUSSION

### *SWAT Model Evaluation*

For improving the calibration of the SWAT model, this study used a variety of input data such as elevation, land use, soil, meteorological/hydrological data, and canal discharge/irrigation. Especially, canal discharge and irrigation are necessary input data to introduce a water path and improve model calibration. In addition, streamflow and water quality are calibrated and validated at Fort Quitman streamflow gauging station for 11 years (2000–2010) and S9 gauging station for 10 years (2002–2011) at the watershed outlet. The calibration and validation periods show different periods between streamflow and water quality data because of the missing data. The SWAT model was simulated for streamflow and water quality including $NO_3$+$NO_2$ and TP. Then, the SWAT model was calibrated and validated. Afterward, the verified SWAT model results were used as the input variables of XGBoost for the prediction of TDS concentrations on the whole river.

A comparison of the observed and simulated monthly streamflow, $NO_3$+$NO_2$, and TP are shown in Figures 4 and 5. SWAT model was calibrated for six years (2000–2005) and validated for five years (2006–2010) of monthly streamflow at Fort Quitman. The average coefficient of determination ($R^2$), Nash–Sutcliffe efficiency (NSE), and the root mean square error for streamflow were 0.60, 0.45, and 3.0 ($m^3$/month). According to the guidelines for SWAT calibration (NSE $\geq$0.5, and $R^2$ $\geq$0.6, Moriasi et al. 2007), the results are found to be satisfactory. As for the water quality, the SWAT model was calibrated for five years (2002–2006) and validated for another five years (2007–2011) of monthly data at the S9 gauging station and the Fort Quitman. The average $R^2$ of $NO_3$+$NO_2$ at the S9 and the Fort Quitman were 0.88 and 0.89, respectively. Also, the average $R^2$ of

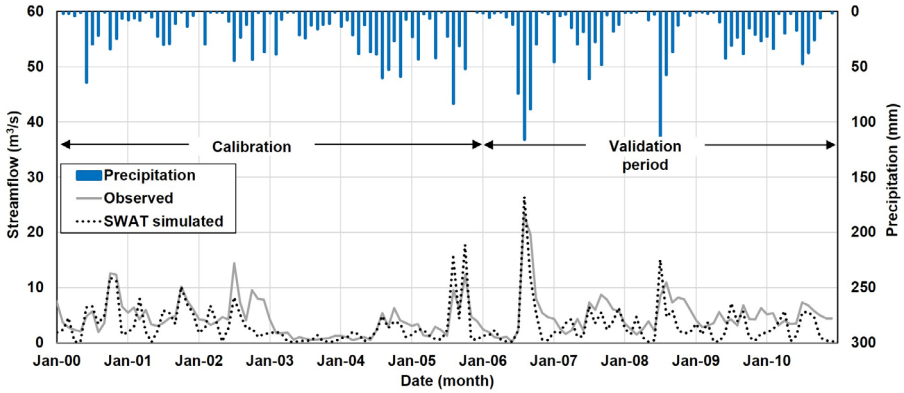


FIGURE 4. The calibration and validation results of observed streamflow vs. simulated SWAT streamflow at Fort Quitman.

TP at the S9 and the Fort Quitman were 0.68 and 0.82, respectively.

### *Evaluation of TDS Performance in Developed XGBoost Algorithm Using Observed Data*

In this study, (1) all 2,505 datasets which consist of nine variables (flow, $NO_3$+$NO_2$, TP, PCP, elevation, main reach length, dominant the topsoil texture, and dominant sublayer soil texture, and observed TDS) per a dataset from 1992 to 2010 at nine stations were randomly divided by 70% (1,753 datasets) and 30% (752 datasets) as training and test datasets, (2) then data were provided to train and test indices using $k$-fold cross-validation with 15 folds, (3) the parameters of XGBoost were optimized using grid-search. The parameters are Max_depth, Booster, Colsample_bytree, and Gamma. The Max_depth is the maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. The Booster is the model of xgboost, that contains low-level routines for training, prediction, and evaluation, which are gbtree, gblinear, or dart. The Colsample_bytree and Gamma are subsample ratio of columns when constructing each tree and minimum loss reduction required to make a further partition on a leaf node of the tree. The optimized parameters of this model were 10.0 for Max_depth, gbtree for Booster, 0.2 for Colsample_bytree, and 2.0 for Gamma, respectively.

Prior to coupling of the XGBoost and the SWAT model, TDS results of XGBoost trained from observed data were evaluated at nine stations. As mentioned before, all independent data based on measurements are flow, $NO_3$+$NO_2$, TP, PCP, elevation, main reach length, dominant topsoil texture, and dominant sublayer soil texture from Yu et al. (2014). From the optimized XGBoost model for algorithm parameters, the TDS concentrations were predicted at all gauging stations. As a result, the average prediction efficiency, expressed as the $R^2$ value, was 0.98 at nine gauging stations (Figure 6). The average $R^2$ for each station were 0.97 at S1, 0.99 at S2, 0.98 at S3, 0.94 at S4, 0.98 at S5, 0.86 at S6, 0.90 at S7, 0.97 at S8, and 0.98 at S9. The average $R^2$ was typically >0.60, which indicates a satisfactory simulation according to Moriasi et al. (2007).

### *Distribution of River Water TDS Using XGBoost and SWAT Model Results*

The TDS distributions on the whole river were predicted by XGBoost based on simulated SWAT results of streamflow, $NO_3$+$NO_2$, and TP of each subwatershed. The yearly spatial distribution maps of river TDS were generated from the monthly results of SWAT (Figure 7). As seen from Figure 7, the distribution of TDS shows higher concentration as going downstream. From three gauging stations, El Paso (inlet), Hudspeth canal, and Fort Quitman (outlet) are selected as watching sites to investigate the progress of increasing TDS (Figure 7).

Figure 8 shows the average TDS concentrations at major three watching sites for 16 years (2000–2015). The average TDS concentrations at three watching sites were 962.6, 1,227.3, and 2,315.4 at El Paso, Hudspeth canal, and Fort Quitman, respectively (Figure 8a). As mentioned before, the TDS increased as the water flows downstream. Figure 8b shows the boxplot of the average TDS at three watching sites. The boxplot is a standardized way of displaying the distribution of data based on five numbers summary such as minimum, first quartile (Q1), median, third quartile (Q3), and maximum with outliers removed. It can also show if the data are symmetrical and how tightly it is grouped or skewed.

To compare the TDS distribution between a wet year and dry year at three watching sites, the

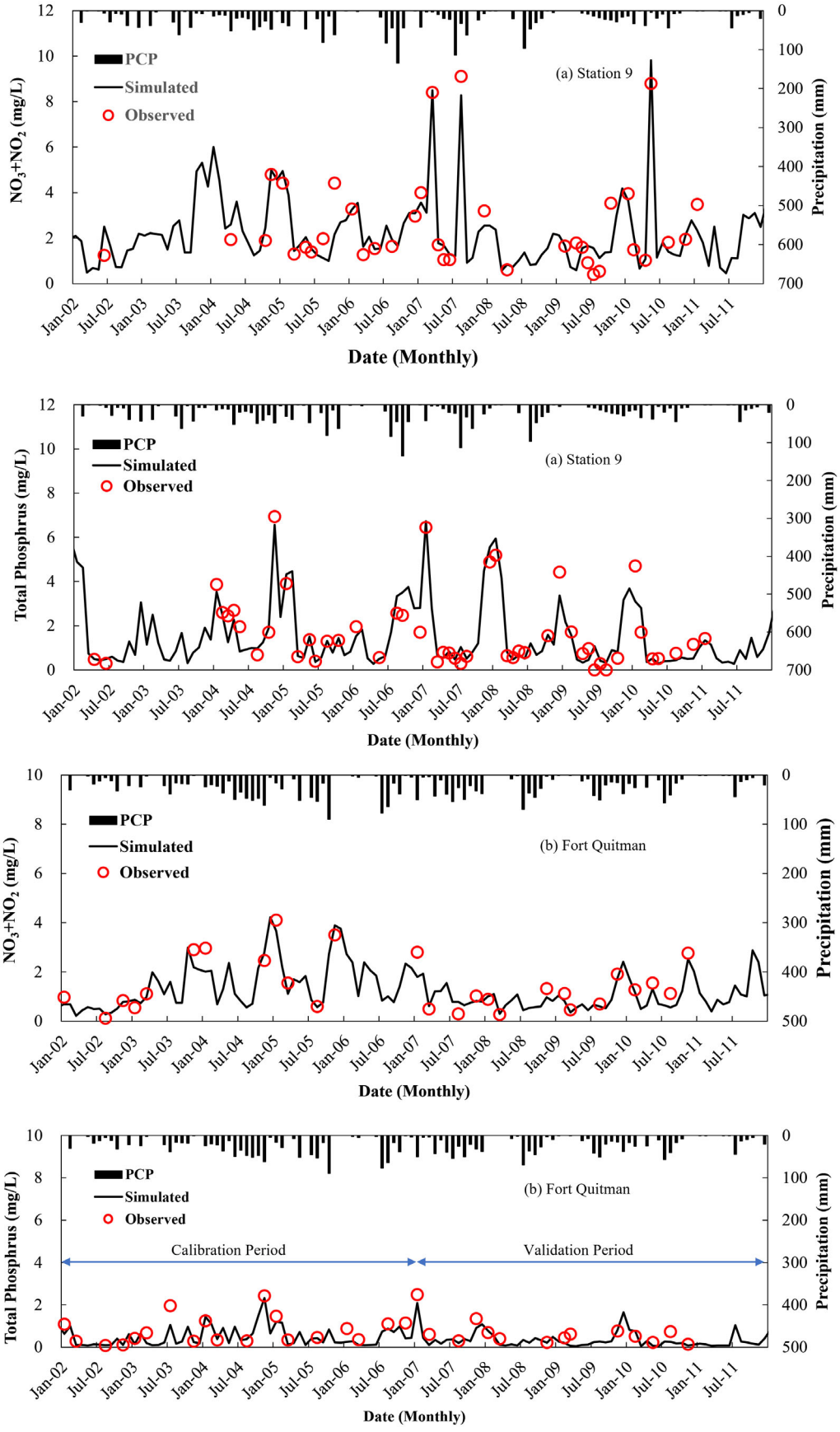


FIGURE 5. The calibration and validation results of observed vs. simulated $NO_3$+$NO_2$ and TP at (a) the S9 gauging station (top two figures) and (b) the Fort Quitman (bottom two figures).

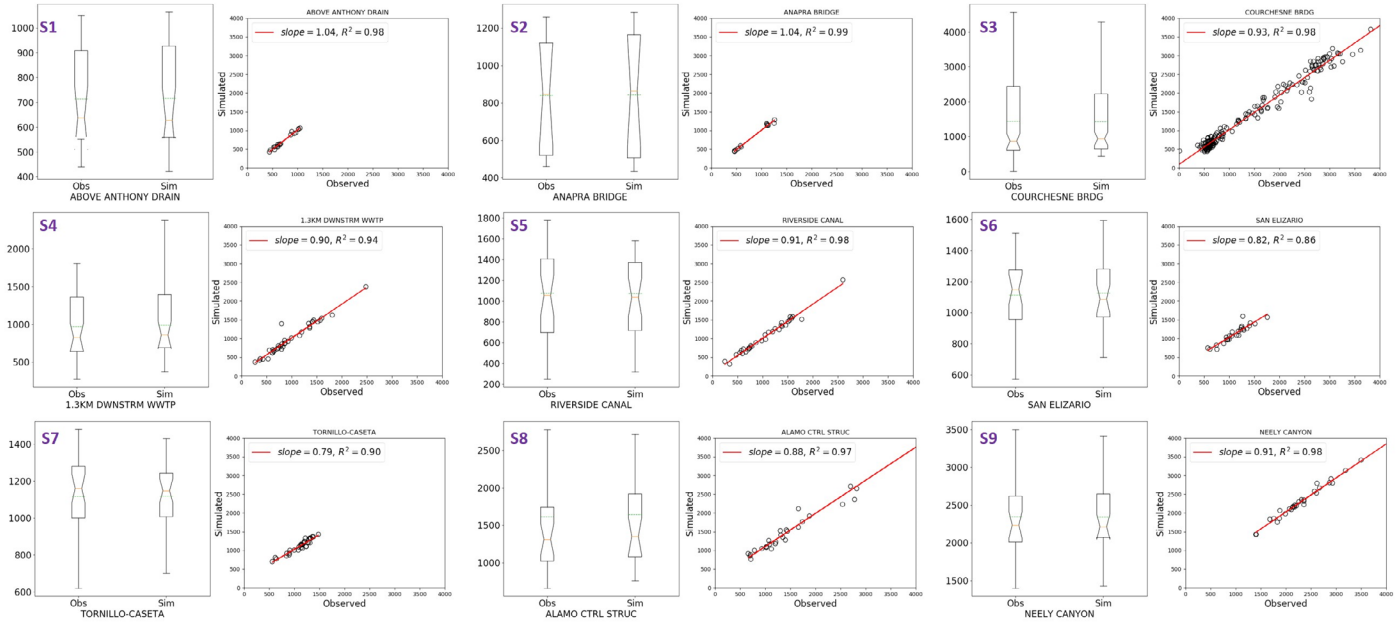


FIGURE 6. The graph results and $R^2$ of the predicted TDS of the XGBoost model vs. observed data at 9 gauging stations.

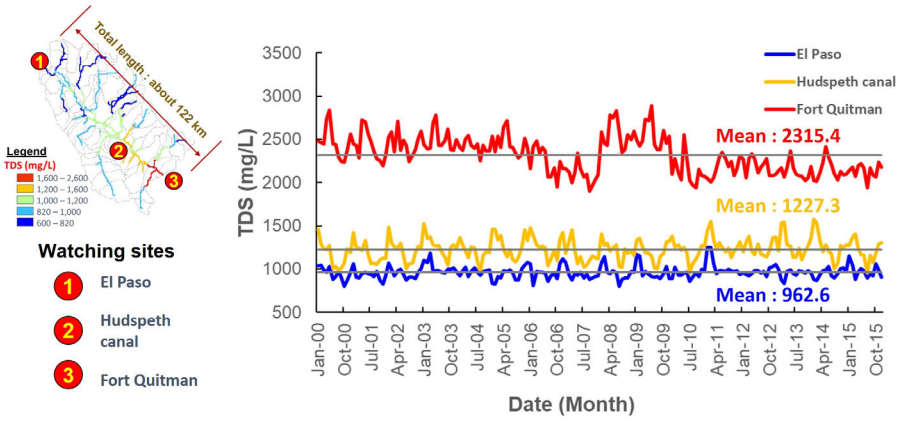


FIGURE 7. The spatiotemporal variations of monthly TDS on the whole river using the XGBoost and SWAT simulated results at three watching sites: (1) El Paso (inlet), (2) Hudspeth canal, and (3) Fort Quitman (outlet).

2007 year as a wet year and 2003 year as a dry year were selected. The PCPs of the wet year and dry year are 371.7 and 138.7 mm, respectively. As shown in Table 2, average TDSs in the dry year were higher than the results of the wet year at all watching sites. On average, the TDSs at El Paso, Hudspeth canal, and Fort Quitman (outlet) in the dry year increased by 8.9%, 4.2%, and 19.7% compared to the results of the wet year, respectively. The groundwater contributions at El Paso, Hudspeth canal, and Fort Quitman, which is the ratio of groundwater discharge divided by total runoff, are 34.6%, 40.9%, and 27.6%. The TDS at Hudspeth Canal in a dry year showed a slight change, which may be attributed to a lower groundwater contribution. Also, each result between wet and dry years was analyzed with the periods divided into nonirrigation (November–March) and irrigation (April–October) seasons. Especially. The results remarkably have shown a different pattern in the dry year. During the dry year, the TDS values at El Paso and Hudspeth canal sites during the nonirrigation season were about 9% and 16% higher than the results for the irrigation season, but on the contrary, the TDS at Fort Quitman during the irrigation season was about 5% higher than that during the nonirrigation season for the study period. Generally, because PCP during the nonirrigation season was less than that during the irrigation season, TDSs at El Paso and Hudspeth canal also have

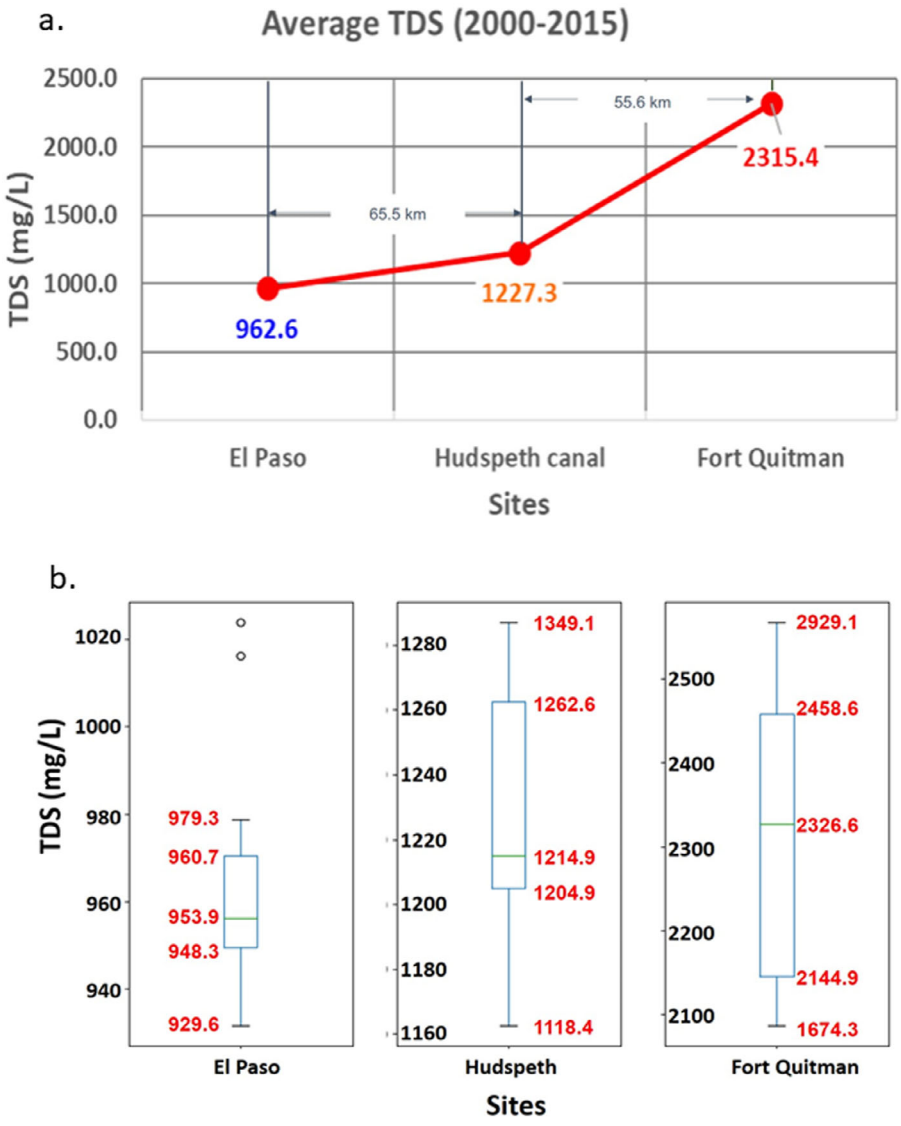


FIGURE 8. The results of (a) predicted average TDS and (b) boxplot of the predicted average TDS from 2000 to 2015 at three watching sites.

negative effects on ramping their concentration up. But, increasing TDS concentration during the irrigation season at Fort Quitman means that irrigation return flow to drains was gathered at Fort Quitman as an outlet. It could be inferred that more saline water was collected than freshwater. By coupling the XGBoost algorithm with SWAT model for complex irrigation systems we can gain a better understanding of the actual dynamics of salinity in the agricultural area.

### *Identification of Causes for High River Water TDS by Soil Texture Scenarios*

This study finally tried to figure out the reasons why the TDS of the river water becomes higher as it flows downstream. The soil salinity level of upland soils is related to soil permeability and irrigation. The salts in these soils were brought in through irrigation water or geological origin. Also, the principal cause of salt accumulation is the low permeability of silty clay loam and silty clay layers, followed by high water tables in a certain area. Soil texture helps determine how much water will be able to pass through the soil, how much water the soil can store, and the ability of salinity to bind to the soil. Hence in this part, changes in TDS can be analyzed and identified by each soil texture at three watching sites. The areas transferring other soil textures at El Paso (inlet), Hudspeth canal, and Fort Quitman (outlet) are 537.1 , 263.8 , and 236.5 $km^2$. The area size will affect the TDS variations by changing soil scenarios.

TABLE 2. Statistical summary of TDS concentrations in the nonirrigation and irrigation periods during wet and dry years at three watching sites.

| | Wet year (2007) | | | | Dry year (2003) | | | |
|---|---|---|---|---|---|---|---|---|
| Period | PCP (mm) | El Paso (TDS, mg/L) | Hudspeth canal (TDS, mg/L) | Fort Quitman (TDS, mg/L) | PCP (mm) | El Paso (TDS, mg/L) | Hudspeth canal (TDS, mg/L) | Fort Quitman (TDS, mg/L) |
| Nonirrigation (November–March) | 78.7 | 946.1 | 1,201.9 | 2,079.8 | 24.2 | 1,069.1 | 1,253.1 | 2,424.5 |
| Irrigation (April–October) | 265.4 | 923.5 | 1,036.2 | 2,087.5 | 114.5 | 978.9 | 1,078.9 | 2,544.0 |
| Mean | 344.1 mm/year | 932.9 | 1,104.8 | 2,084.3 | 138.7 mm/year | 1,016.2 | 1,151.0 | 2,494.5 |

Note: PCP, precipitation.

As shown in Table 3, TDS at each site was predicted by changing assumed soil textures such as very cobbly, very fine sandy loam, loam, silt loam, and silty clay loam from the original texture. At each site, original soil textures are very cobbly for El Paso, very fine sandy loam at Hudspeth canal, and silt loam at Fort Quitman. During the simulated period (2000–2015), the average TDS at El Paso showed the lowest concentration of 962.6 mg/L in having original soil texture as very cobbly. On the other hand, TDS with silt loam has the highest concentration of 1,053.6 mg/L. The TDS increased by 9.4% compared to TDS of the original soil texture. At the Hudspeth canal, TDS changes ranged from −4.4% to +2.2% compared to the original soil texture as very fine sandy loam. It also showed that TDS with silt loam has the highest concentration and TDS with very cobbly has the lowest concentration. In Fort Quitman at the watershed outlet, the original soil texture was silt loam. So, TDS changes decreased from 0.2% to 5.5% in comparison to the original soil texture.

As for the TDS changes, it could come from soil structure and physical characteristics. Because soil is composed of small particles, silt soils can hold more water and are slower to drain than coarse-textured soils. Smaller particles can pack closely together, block the spaces between particles, and prevent water from passing through. Whereas sand particles are larger and therefore, have larger pore spaces for water to pass through. Under normal irrigation practices, sandy soils will naturally be able to flush more water through the root zone than clay soils. The end result is that sandy soils can withstand higher salinity irrigation water because more dissolved salts will be removed from the root zone by leaching. For soil salinity mitigation from these results, a soil reclamation plan can be implemented. Especially, if the soil to be reclaimed has a heavy texture (i.e., silt or clay soils), the mixing of sand in an appropriate quantity can change the soil texture permanently; the soil becomes more permeable and is easier to reclaim. Changing the soil texture is a difficult and costly task, though where sand is readily available, such as in a desert, this practice can be accomplished more easily (Shahid et al. 2018).

## DISCUSSIONS

Even though the streamflow and water quality simulation from the results has relatively good

TABLE 3. Summary of changed average TDS with soil scenarios by assuming other soil texture compared to TDS concentration of the original soil for 16 years (2000–2015).

| | Sites | | | | | |
|---|---|---|---|---|---|---|
| | El Paso (very cobbly) | | Hudspeth canal (very fine sandy loam) | | Fort Quitman (silt loam) | |
| Soil texture | Concentration (mg/L) | Change (%)[1] | Concentration (mg/L) | Change (%)[1] | Concentration (mg/L) | Change (%)[1] |
| Very cobbly | 962.6 | Original soil | 1,173.1 | −4.4 | 2,189.0 | −5.5 |
| Very fine sandy loam | 1,046.2 | +8.7 | 1,227.3 | Original soil | 2,280.0 | −1.5 |
| Loam | 1,011.0 | +5.0 | 1,235.7 | +0.7 | 2,288.3 | −1.2 |
| Silt loam | 1,053.6 | +9.4 | 1,254.0 | +2.2 | 2,315.4 | Original soil |
| Silty clay loam | 1,052.0 | +9.3 | 1,231.0 | +0.3 | 2,311.2 | −0.2 |

[1](TDS concentration of the original soil − TDS concentration of the other soil)/(TDS concentration of the original soil) × 100.

accuracy according to Moriasi et al. (2007), some limitations should be analyzed. The study area is the semi-arid region and the river in the area is almost kept dry. Also, most of the surface water has been artificially consumed as agricultural water into each canal. So, the streamflow in this area cannot be defined as a natural flow. Due to the artificial system, the model has some limitations for the prediction of streamflow and led to lower model performance.

In the general process of fitting the regression method, when one independent variable is nearly a combination of other independent variables, there will affect parameter estimates. This problem is called multicollinearity. While multicollinearity is not a violation of the assumptions of regression, it may cause serious difficulties (Neter and Wasserman 1989; Lin 2008): (1) variances of parameter estimates may be unreasonably large, (2) parameter estimates may not be significant, and (3) a parameter estimate may have a significant difference from what is expected, and so on. For solving or alleviating this problem in certain regression, the best way is to drop redundant variables from this model directly, that is to try to avoid it by not including redundant variables in the regression method (Bowerman and O'Connell 1993; Lin 2008). All observed data of this study could be selected as independent variables because these data completely did not affect each other. Thus, all observed data such as flow, $NO_3$+$NO_2$, TP, TDS as well as PCP, elevation, main reach length, the dominant top layer soil texture, and the dominant sublayer soil texture to consider geographical characteristics are finally selected as independent variables. The geographical characteristics were considered using the recommendations of Yu et al. (2014).

As for the performance of the coupling model, it could have remarkable strength in terms of tracking vulnerable sites using the spatial distribution of the river salinity. As seen from Figure 7, salinity got worse downstream than upstream. In particular, the salinity became more severe after passing the Hudspeth canal. On average, the salinity at the Hudspeth canal increased 127.5% than El Paso. At this site, as shown in Figure 1, the confluence of irrigation network from both Mexico and the U.S. may have played an important role in elevated salinity, which attributes to nonpoint sources and wastewater discharge from Mexico. Therefore, appropriate salinity control measures should be taken to reduce the impacts of the salinity downstream from this location. This coupling model results of spatial distribution on river salinity could help us track high salinity spots and develop strategies for salinity mitigation and management.

## SUMMARY AND CONCLUSIONS

This study successfully tested a new approach for assessing salinity distribution within a watershed by coupling the XGBoost algorithm and the SWAT model and identifying the reasons why the river water salinity gets worse at three watching sites from different perspectives, which could provide guidelines for water and soil salinity management. The XGBoost was developed and trained by using observed hydrology, water quality, and geographic data at nine gauging stations. The SWAT model was calibrated and validated for streamflow and water quality to be used as input variables of the XGBoost for monthly TDS prediction. Then, the spatial distribution of the TDS on the river system within the watershed was predicted using nine variables selected with the developed XGBoost algorithm. Finally, TDS changes at each watching site were analyzed by changing assumed soil textures.

By improving both bias and variance for TDS estimation, the method used in this study demonstrated advantages over other methods such as using only the model and simple regression. Nevertheless, there were not enough data at most stations. The number of the data at each station was <150 for 14 years and there were also a lot of missing water quality data. Therefore, most data were used to train the XGBoost model and it might have caused an overfitting problem in predicting TDS. To solve such a problem, additional data collection for TDS and other water quality parameters is recommended.

In addition, due to missing observed data, the SWAT model was not fully calibrated for streamflow and water quality, especially in the irrigation network (canals and drains). And if the SWAT model could consider other water quality components such as TSS, chloride, and total organic carbon, this method can become a more useful tool for the study of salinity dynamics. For future research, it is recommended to focus on the mitigation and adaptation measures to reduce river salinity, safeguarding agricultural production and municipal water supplies.

## FUNDING INFORMATION

This work was supported in part by USDA grant (Project No. 2017-68007-26318) through the National Institute of Food and Agriculture's Agriculture and Food Research Initiative, Water for Agricultural Challenge Area.

ACKNOWLEDGMENT

The authors appreciate the support in part by the National Institute of Food and Agriculture (NIFA) Hatch Project of the U.S. Department of Agriculture.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## AUTHOR CONTRIBUTIONS

**Chunggil Jung:** Conceptualization; Formal analysis; Methodology; Software; Writing-original draft. **Sora Ahn:** Formal analysis; Methodology; Writing-review & editing. **Zhuping Sheng:** Conceptualization; Formal analysis; Funding acquisition; Methodology; Project administration; Supervision; Writing-review & editing. **Essayas K. Ayana:** Data curation; Writing-review & editing. **Raghavan Srinivasan:** Data curation; Funding acquisition; Project administration; Supervision; Writing-review & editing. **Dhanesh Yeganantham:** Data curation; Writing-review & editing.

## LITERATURE CITED

Abudu, S., J.P. King, and Z. Sheng. 2012. "Comparison of the Performance of Statistical Models in Forecasting Monthly Total Dissolved Solids in the Rio Grande." *Journal of the American Water Resources Association* 48 (1): 10–23. https://doi.org/10.1111/j.1752-1688.2011.00587.x.

Adhikari, U., A.P. Nejadhashemi, and S.A. Wonznicki. 2015. "Climate Change and Eastern Africa: A Review of Impact on Major Crops." *Food and Energy Security* 4 (2): 110–32. https://doi.org/10.1002/fes3.61.

Arnold, J.G., R. Srinivasan, R.S. Muttiah, and J.R. Williams. 1998. "Large Area Hydrologic Modeling and Assessment Part I: Model Development 1." *Journal of the American Water Resources Association* 34 (1): 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Bowerman, B.L., R.T. O'Connell, and T. Richard. 1993. *Forecasting and Time Series: An Applied Approach*. Belmont, CA: Wadsworth.

Chen, C., W. He, H. Zhou, Y. Xue, and M. Zhu. 2020. "A Comparative Study among Machine Learning and Numerical Models for Simulating Groundwater Dynamics in the Heihe River Basin, Northwestern China." *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-60698-9.

Cisty, M., and V. Soldanova. 2018. "Flow Prediction versus Flow Simulation using Machine Learning Algorithms." In *Machine Learning and Data Mining in Pattern Recognition*, edited by P. Perner, 14th International Conference, MLDM 2018, New York, NY, July 15–19, 2018, Proceedings, Part II: 369–82.

Doremus, D., and G. Lewis. 2008. "Rio Grande Salinity Management — A Real Possibility?" *Southwest Hydrology* 7 (2): 24–25.

Eissa, M.A., J.M. Thomas, G. Pohll, O. Shouakar-Stash, R.L. Hershey, and M. Dawoud. 2016. "Groundwater Recharge and Salinization in the Arid Coastal Plain Aquifer of the Wadi Watir Delta, Sinai, Egypt." *Applied Geochmi* 71: 48–62.

El-Bihery, M.A. 2009. "Groundwater Flow Modeling of Quaternary Aquifer Ras Sudr, Egypt." *Environmental Geology* 58 (5): 1095–105.

FAO. 2011. *The State of the World's Land and Water Resources for Food and Agriculture*. Rome and Earthscan, New York: The Food and Agriculture Organization of the United Nations.

Ghorbani, K., A. Wayayok, A. Fikri, and M. Abbaszadeh. 2017. "Investigation of Salinity Consequences Resulting from Drainage Systems using Numerical Models." *Journal of Irrigation and Drainage Engineering* 143 (5). https://doi.org/10.1061/(ASCE)IR.1943-4774.0001100.

Hanson, B., J.W. Hopmans, and J. Simunek. 2008. "Leaching with Subsurface Drip Irrigation under Saline, Shallow Groundwater Conditions." *Vadose Zone Journal* 7 (2): 810–18.

Heywood, C.E., and R.F. Yager. 2003. "Simulated Ground-Water Flow in the Hueco Bolson, an Alluvial-Basin Aquifer System Near El Paso, Texas." *U.S. Geological Survey, Water Resources Investigations Report 02-4108*, p. 73.

Hogan, J.F., F.S. Phillips, K. Mills, J.M.H. Hendricks, J. Ruiz, J. Chesley, and Y. Asmerome. 2007. "Geologic Origins of Salinization in a Semi-arid River: The Role of Sedimentary Basin Brines." *Geology* 35 (12): 1063–66.

Hutchison, W.R. 2004. "Hueco Bolson Groundwater Conditions and Management in the El Paso Area." El Paso Water Utilities, Hydrogeology Report 04-01, March 2004.

Ibrahimi, M.K., T. Miyazaki, T. Nishimura, and H. Imoto. 2014. "Contribution of Shallow Groundwater Rapid Fluctuation to Soil Salinization under Arid and Semiarid Climate." *Arabian Journal of Geosciences* 7 (9): 3901–11.

Kanzari, S., M. Hachicha, R. Bouhlila, and J. Battle-Sales. 2012. "Characterization and Modeling of Water Movement and Salts Transfer in a Semi-arid Region of Tunisia (Bou Hajla, Kairouan) — Salinization Risk of Soils and Aquifers." *Computers and Electronics in Agriculture* 86: 34–42.

Karandish, F., and J. Simunek. 2016. "A Comparison of Numerical and Machine-learning Modeling of Soil Water Content with Limited Input Data." *Journal of Hydrology* 543: 892–909.

Lamorski, K., T. Pastuszka, J. Krzyszczak, C. Stawinski, and B.W. Walczak. 2013. "Soil Water Dynamic Modelling Using Physical and Support Vector Machine Methods." *Vadose Zone Journal* 12 (4). https://doi.org/10.2136/vzj2013.05.0085.

Lin, F.J. 2008. "Solving Multicollinearity in the Process of Fitting Regression Model Using the Nested Estimate Procedure." *Quality & Quantity* 42: 417–26.

Mehta, P., M. Bukov, C.H. Wang, A.G.R. Day, C. Richardson, C.K. Fisher, and D.J. Schwab. 2019. "A High-Bias, Low-Variance Introduction to Machine Learning for Physicists." *Physics Reports* 810: 1–124.

Montegomery & Associates, and W.R. Hutchison. 2016. "Groundwater Flow and Transport Model for Hueco Bolson Aquifer El Paso and Hudspeth Counties, Texas." Prepared for El Paso Water Utilities; EPW Hydrogeology Report — Volume 1, December 2016.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith. 2007. "Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations." *Transactions of the American Society of Agricultural and Biological Engineers* 50 (3): 885–900. https://doi.org/10.13031/2013.23153.

Moyer, D.L., S.K. Anderholm, J.F. Hogan, F.M. Phillips, B.J. Hibbs, J.C. Witcher, A.M. Matherne, and S.E. Falk. 2009. "Knowledge and Understanding of Dissolved Solids in the Rio Grande — San Acacia, New Mexico, to Fort Quitman, Texas, and Proposed Plan for Future Studies and Monitoring." *U.S. Geological Survey OpenFile Report 2013-1190*.

Nalenz, M., and M. Villani. 2018. "Tree Ensembles with Rule Structured Horseshoe Regularization." *The Annals of Applied Statistics* 12 (4): 2379–408.

Neitsch, S.L., J.G. Arnold, J.R. Kiniry, and J.R. Williams. 2009. *Soil and Water Assessment Tool Theoretical Documentation Version 2009*. College Station, TX: Texas Water Resources Institute.

Neter, J., W. Wasserman, and M.H. Kutner. 1989. *Applied Linear Regression Models*. Homewood, IL: Richard D. Irwin.

Park, J.Y., S.R. Ahn, S.J. Hwang, C.H. Jang, G.A. Park, and S.J. Kim. 2014. "Evaluation of MODIS NDVI and LST for Indicating Soil Moisture of Forest Areas Based on SWAT Modeling." *Paddy and Water Environment* 12 (Suppl. 1): 77–88. https://doi.org/10.1007/s10333-014-0425-3.

Park, J.Y., M.J. Park, S.R. Ahn, G.A. Park, J.E. Yi, G.S. Kim, R. Srinivasan, and S.J. Kim. 2011. "Assessment of Future Climate Change Impacts on Water Quantity and Quality for a Mountainous Dam Watershed Using SWAT." *Transactions of the American Society of Agricultural and Biological Engineers* 54 (5): 1725–37. https://doi.org/10.13031/2013.39843.

Phillips, F.M., J.F. Hogan, S.K. Mills, and J.M.H. Hendrickx. 2003. "Environmental Tracers Applied to Quantifying Causes of Salinity in Arid Region Rivers: Preliminary Results from the Rio Grande, Southwestern USA." In *Water Resources Perspectives: Evaluation, Management, and Policy*, edited by A.S. Alsharhan, and W.W. Wood, 327–34. New York: Elsevier.

Postel, S. 1999. *Pillar of Sand: Can the Irrigation Miracle Last?*. New York, London: W. W. Norton & Company Ltd.

Punnoose, R. and P. Ajit. 2016. "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms." *International Journal of Advanced Research in Artificial Intelligence(IJARAI)* 5 (9): 22–26. http://doi.org/10.14569/IJARAI.2016.050904

Qin, Z., D.B. Myers, C.J. Ransom, N.R. Kitchen, S. Liang, J.J. Camberato, P.R. Carter *et al*. 2018. "Application of Machine Learning Methodologies for Predicting Corn Economic Optimal Nitrogen Rate." *Agronomy Journal* 110 (6): 2596–607.

Ragab, R., and C. Prudhomme. 2002. "Climate Change and Water Resources Management in Arid and Semi-arid Regions: Prospective and Challenges for the 21st Century." *Biosystems Engineering* 81 (1): 3–34. https://doi.org/10.1006/bioe.2001.0013 .

Schwabe, K.A., I. Kan, and K.C. Knapp. 2006. "Drain Water Management for Salinity Mitigation in Irrigated Agriculture." *American Journal of Agricultural Economics* 88: 133–49. https://doi.org/10.1111/j.1467-8276.2006.00843.x.

Shahid, S.A., M. Zaman, and L. Heng. 2018. *Guideline for Salinity Assessment, Mitigation and Adaptation Using Nuclear and Related Techniques*. Cham, Switzerland: Springer.

Sheng, Z. 2013. "Impacts of Groundwater Pumping and Climate Variability on Groundwater Availability in the Rio Grande Basin." *Ecosphere* 4 (1): 1–25.

Sheng, Z., and J. Devere. 2005. "Understanding and Managing the Stressed Mexico-USA Transboundary Hueco Bolson Aquifer in the El Paso del Norte Region as a Complex System." *Journal of Hydrogeology* 13 (5–6): 813–25.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15: 1929–58.

Tuv, E., A. Borisov, G. Runger, and K. Torkkola. 2009. "Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination." *Journal of Machine Learning Research* 10: 1341–66.

Umali, D.L. 1993. "Irrigation-Induced Salinity: A growing Problem for Development and Environment." World Bank Technical Paper No. 215, World Bank, Washington, D.C.

USACE. 2011. *Alternatives Analysis for the Rio Grande Salinity Management Program*. Washington, D.C.: U.S. Army Corps of Engineers.

Vandenberghe, V., W. Bauwens, and P.A. Vanrolleghem. 2007. *Evaluation of Uncertainty Propagation into River Water Quality Predictions to Guide Future Monitoring Campaigns*. Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Vermeulen, D., and A.V. Niekerk. 2017. "Machine Learning Performance for Predicting Soil Salinity Using Different Combinations of Geomorphometric Covariates." *Geoderma* 299: 1–12.

White, E. 2015. "Predicting Unimpaired Flow in Ungauged Basins: 'Random Forests' Applied to California Streams." Master of Science thesis, University of Kentucky.

Wu, W., C. Zucca, A. Muhaimeed, W. Al-Shafie, A.M.F. Al-Quraishi, V. Nangia, M. Zhu, and G. Liu. 2018. "Soil Salinity Prediction and Mapping by Machine Learning Regression in Central Mesopotamia, Iraq." *Land Degradation & Development* 29 (11): 4005–14.

Yu, J., Y. Li, G. Han, D. Zhou, Y. Fu, B. Guan, G. Wang, K. Ning, H. Wu, and J. Wang. 2014. ""The Spatial Distribution Characteristics of Soil Salinity in Coastal Zone of the Yellow River Delta." *Environmental Earth Sciences* 72 (2): 589–99.

Zhang, D., L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si. 2018. "A Data-driven Design for Fault Detection of Wind Turbines using Random Forests and Xgboost." *IEEE Access* 6: 21020–21031.