Chapter 25
Salinity Prediction in Coastal Aquifers of the Vietnamese Mekong River Delta Using Innovative Machine Learning Algorithms

Dang An Tran, Ha Nam Thang, Dieu Tien Bui, and Vuong Trong Kha

Abstract Groundwater salinization is a severe issue, causing various problems to human health, agriculture, ecosystems, and infrastructure in many coastal regions across the world. However, this phenomenon is difficult to predict with high accuracy. In this study, we propose and verify a new artificial intelligence approach for predicting groundwater salinity and identifying the main factors of salinization. The coastal aquifers of the Mekong River Delta (Vietnam) were selected to test the new approach. In the proposed approach, Extreme Gradient Boosting (XGB) was used to build a groundwater salinity model, and Genetic Optimization (GO) was employed to optimize the model parameters. Gaussian Processes (GP) and Random Forests (RF) were also used as a benchmark for the model comparison. For this regard, a groundwater salinity database with 215 groundwater samples and 20 driven factors related to hydrology, geology, geography, and anthropogenic activities was prepared. Performance of the models was assessed using Correlation Coefficient (r), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE). The results show that the proposed GO-XGB model yields high performance both on the training dataset (r = 0.999, RMSE = 18.450, MAPE = 2.070, and MAE = 4.864) and the validation dataset (r = 0.787, RMSE = 141.042, MAPE = 87.250, and MAE = 74.993). The proposed GO-XGB model

D. A. Tran (✉)
Faculty of Water Resources Engineering, Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam
e-mail: antd@tlu.edu.vn

H. N. Thang
Faculty of Fisheries, University of Agriculture and Forestry, Hue University, Hue 530000, Vietnam
e-mail: hanamthang@hueuni.edu.vn

D. T. Bui
Department of Business and IT, USN School of Business, Campus Bø (1-324), Bø, Norway
e-mail: dieu.t.bui@usn.no

V. T. Kha
Department of Mine Surveying, Hanoi University of Mining and Geology, Hanoi, Vietnam

P. L. Vo et al. (eds.), *Advances in Research on Water Resources and Environmental Systems*, Environmental Science and Engineering,
https://doi.org/10.1007/978-3-031-17808-5_25

performed better predictive result compared to the benchmark, GP, and RF. Among the 20 factors, groundwater level, vertical hydraulic conductivity, lithology, extraction capacity, horizontal hydraulic conductivity, distance to saline sources, and well density are the most important factors to groundwater salinization prediction.

Keywords Groundwater salinization · GO-XGB model · Coastal aquifers · Mekong River Delta · Vietnam

25.1 Introduction

Groundwater is identified as the primary source of water for about two billion people and accounts for 33% of the total water withdrawal worldwide (Famiglietti 2014). It is a crucial freshwater resource for domestic uses, industrial development, and irrigational activities (Mohanty and Rao 2019; Behera et al. 2019; Kaur et al. 2020). However, groundwater resource is highly vulnerable to human activities (Ma et al. 2019a; Brouwer et al. 2018; Graaf et al. 2019) and natural variation (Kagabu et al. 2020; Giambastiani et al. 2018), especially in coastal regions where are facing groundwater overexploitation, seawater intrusion, climate change, and sea-level rise (Ferguson and Gleeson 2012). In such regions, groundwater is likely to increase in salinity due to paleo-seawater intrusion (Delsman et al. 2014), modern seawater intrusion (Han and Currell 2018), leaking brines from oil fields and irrigation activities (Paine 2003). High salt concentrations in groundwater may cause various environmental and health issues. For example, high salinity in irrigated water may cause physiological drought and reduce crop yield (Nishanthiny et al. 2010). High salt in drinking water increases the risk of hypertension (Vineis et al. 2011), coronary heart disease (Park and Kwock 2015), and chronic kidney disease; therefore, assessing groundwater quality, especially the salinization level, is crucial to protect the environment and human health (Melloul and Goldenberg 1997; Guhl et al. 2006; Gallardo and Marui 2007; Carretero et al. 2013; Larsen et al. 2017).

For last several decades, mathematical model has been used widely in prediction of groundwater dynamics and seawater intrusion into coastal aquifers (Lal and Datta 2019; Abdelhamid et al. 2016; Mahmoodzadeh and Karamouz 2019; Stein et al. 2019; Voss and Souza 1987). However, mathematical groundwater modelling requires expert knowledge about the physical characteristics of hydrogeological system, governing process, various types of input data (i.e., topography, soil properties, geology, initial and boundary conditions, hydrological and climate data, etc.) while the accuracy of the model simulation depends on reliable model input parameters (Lal and Datta 2019; Kim and Yang 2018). Meanwhile, machine learning is a data-driven model with little requirement about the physical process, and it could provide an accurate prediction (Sun et al. 2016; Yadav et al. 2018). Therefore, machine learning has been considered as an alternative, i.e., Genetics algorithm (Sreekanth and Datta 2010), artificial neural networks (Banerjee et al. 2011), multi-objective optimization (Javadi et al. 2015), multivariate adaptive regression spline

(Roy Dilip and Datta 2017), support vector regression (Lal and Datta 2019; Isazadeh et al. 2017; Nadiri et al. 2018), ensemble multiadaptive boosting logistic regression (Rizeei et al. 2019), and Gaussian Process Regression (Yadav et al. 2018; Kopsiaftis et al. 2019), and hybrid computational intelligence models (Pham et al. 2019a; Chen et al. 2019). A common conclusion from the above works is that machine learning is a highly flexible tool with the ability to handle complex non-linear relationships between groundwater salinity and influencing factors (Naghibi et al. 2015; Ransom et al. 2017; Sajedi-Hosseini et al. 2018). Nonetheless, no studies have figured out which are the most important factors influencing on groundwater salinity in coastal areas, while the rapid development in the field of computer science has introduced more superior methods.

Inspite of many advantages of applying machine learning in predicting environmental issues, this approach has some limitations such as lacking good data, deterministic problems, and misapplication. Especially, the predictive results mainly based on statistical relationship instead of performing directly physical processes like numerical models therefore it requires in-depth understanding between target variable and independent variables to improve reliability and accuracy of the ML models. In this research, therefore, we propose and validate a new artificial intelligence approach, which is based on Extreme Gradient Boosting (XGB) and Genetic Optimization (GO), named as GO-XGB, for predicting groundwater salinity in the coastal aquifers of the Mekong River Delta (Vietnam). To the best of our knowledge, this is the first time that GO-XGB is considered for groundwater salinity modelling. We also compare and discuss the performance of our models and traditional models such as random forests and Gaussian processes to understand if this approach adds value to the field of groundwater salinity prediction. Besides, the role of various influencing factors in aquifer salinization is assessed. The proposed models were tested using groundwater salinity data and its controlling factors in the multi-aquifers in the Mekong Delta, Vietnam.

25.2 Background of the Machine Learning Algorithms Used

In this section, we first review two traditional machine learning models which are already applied to predict groundwater salinity, namely random forests, and Gaussian processes. We then introduce the idea of the combination of Extreme Gradient Boosting and Genetic Optimization to form a new hybrid algorithm. The performance of the two traditional models is then considered as benchmarks to assess our model.

25.2.1 Gaussian Processes

Gaussian processes (GP) are a type of supervised learning for both regression and classification problems (Kopsiaftis et al. 2019; Rasmussen et al. 2003; Hall et al. 2012; Azimi et al. 2018). The principal idea of Gaussian processes is that in the input space x = [x_1, ..., x_n] T, every point is associated with a random variable, so as the joint distribution of them can be modelled as a multivariate Gaussian and a function (called f) can be modelled using an infinite multivariate Gaussian distribution (Ma et al. 2019b). Similarly, if we have a salinity dataset M = ([X_i, y_i], i = 1, 2, …, m) with $X_i \in R_n$ is a matrix of m input variables with n observation, whereas $y_i \in R$ is an output variable (Cl^- concentration in groundwater). A GP regression model formulates the relation of the input and output variables as following equation (Rasmussen et al. 2003; Hoa et al. 2019):

$$y(x) = \sum_{i=1}^{n} \alpha_i K(X_i, X) \tag{25.1}$$

where α_i is the weight and K is the Radial Basis kernel function (RBF) (Eq. 25.2) (Park and Sandberg 1991; Scholkopf et al. 1997).

$$K(X_i, X) = \beta \times e^{-\sum_{i=1}^{m}\left[\frac{(X_i^m - X_i^m)^2}{2\sigma^2}\right]} \tag{25.2}$$

where β is the scaling factor and σ is the kernel parameter.

The performance of the GP model is dependent on the parameters β and weights α_i and they could be automatically turned and optimized through maximizing the marginal likelihood (Rasmussen et al. 2003).

25.2.2 Random Forests

A random forest (RF) is a method for both classification and regression based on the ensemble of decision trees (Breiman 2001). A decision tree is a top-down tree-like structure, in which each non-leaf node is a test, each branch is an outcome of the test, and each leaf node is a decision. Regression with a single decision tree may result in the problem of overfitting (high variance) and is dependent on the distribution of training sets. A large number of decorrelated decision trees can form a random forest which then can reduce the variance and boost model performance (Criminisi 2011). The procedure developing RFs is as follows: (1) n random subsets (called "bootstrapped subsets") are sampled from a training dataset based on a random selection of features of the dataset. A subset may contain overlapped data in other

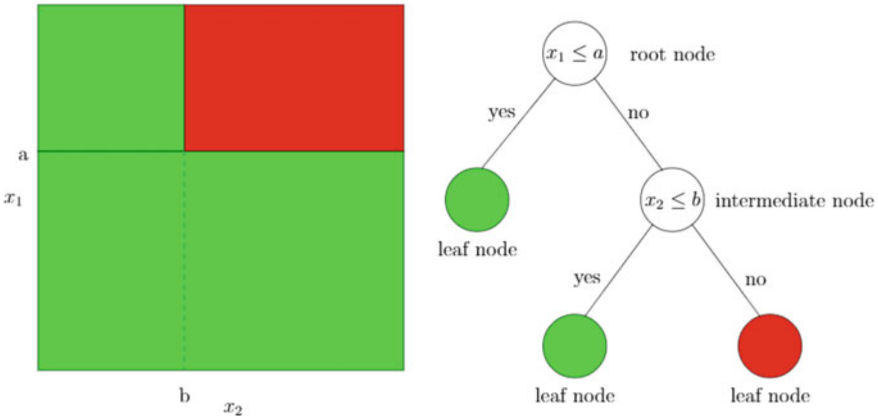

Fig. 25.1 Example of the partitions left and classification tree structure right with two classes coloured in green and red

subsets; (2) n decision trees are built using these n bootstrapped subsets (Fig. 25.1). The number of trees n is decided using either cross-validation or out-of-bag (OOB) error methods. A detailed description of the statistical formulation of RF can be found in Breiman (2001).

25.2.3 Extreme Gradient Boosting

Similar to the random forest, an Extreme Gradient Boosting (XGB) is an ensemble-machine learning algorithm that is based on decision trees (weak learners) (Friedman 2001). However, a boosting model constructs the "forest" of decision trees sequentially, or one decision tree can be constructed based on learning experience inherited from previous trees (Chen and Guestrin 2016; Johnson et al. 2018). The second tree focuses on the cases in which the first tree gives a poor prediction, and this learning process is repeated many times, so as the combination of these trees can better capture the relationship between predictands and predictors. Gradient Boosting is a form of boosting models in which poor prediction cases are assessed if they contribute to minimize the overall lost function (also called the prediction error) (Lim and Chi 2019). A case can be considered as highly valuable if the adding decision tree built for this case can reduce the prediction error significantly while no change in the error implicates a no value case; thus, only useful decision trees are kept. This may give XGB models advantages in complex problems like quantifying saline concentration in groundwater since data measurement in the underground environment may contain many special cases. It is also worth to notice that the learning efficiency of each machine learning algorithm is controlled by its model parameters, and in the case of the XGB model, they include three groups: tree-specific, boosting, and miscellaneous parameters. Selection of these model parameters is a challenging task and depends on user experience while this process does not always return in an optimum set of parameters. Thus, we propose to use a genetic algorithm to automatically search in parameter spaces to improve the accuracy of numerical forecasts.

25.2.4 Genetic Algorithm

Genetic Algorithm takes the idea from the Darwinian theory of natural selection to evolve solutions by utilizing computer capacity to tune model parameters as an alternative to manual efforts (Forrest 1993). The most crucial concept of GA is the chromosome which consists of model parameters to define a solution (called individual) (Jennings et al. 2019). A certain number of individuals then forms a population. In the lower level, each chromosome consists of some genes which are often denoted as 0 s or 1 s ($X \equiv (x_1, x_2, \ldots, x_n)$, $x_k \in [0.0, 1.0]\ \forall\ k$). Each individual is evaluated by its fitness value, a result of a fitness function.

The basic operation performed during the training of XGB based model is as following steps: (1) A number of individuals are initialized to form a population, (2) individuals with the best fitness values are selected to generate a mating pool, (3) from the mating pool, either sequential or random selection methods select parents, and (4) several operators called crossover and mutation are then applied to each pair of parents to generate their offspring. This process keeps high-quality individuals to create more individuals, so as it evolves solutions to obtain the desired solutions.

25.3 Study Area and Data

25.3.1 Description of the Study Area

The study area, Soc Trang province, is in the coastal area of the Mekong River Delta. The study area covers an area of 3,312 km^2 with an elevation ranging from 0.5 to 2.5 m above the mean sea level (Fig. 25.2). The province is bordered by the Hau River (one main branch of the Mekong River) to the Northwest and the Vietnamese East Sea (South China Seas) to the Southwest. Since this area has a dense river system connected to the sea, the hydrological regime in the study area is complex and strongly influenced by the flow regime of the Mekong River and tidal fluctuation.

The study area is in a tropical monsoon climate region with two distinct seasons, the dry season from May to November and the rainy season from December to April (in the following year). The annual average rainfall is about 1772 mm with substantial seasonal variation. About 85% of the annual rainfall occurs during the rainy season. The study area has recognized as one of the most vulnerable regions to climate change and sea-level rise in the world.

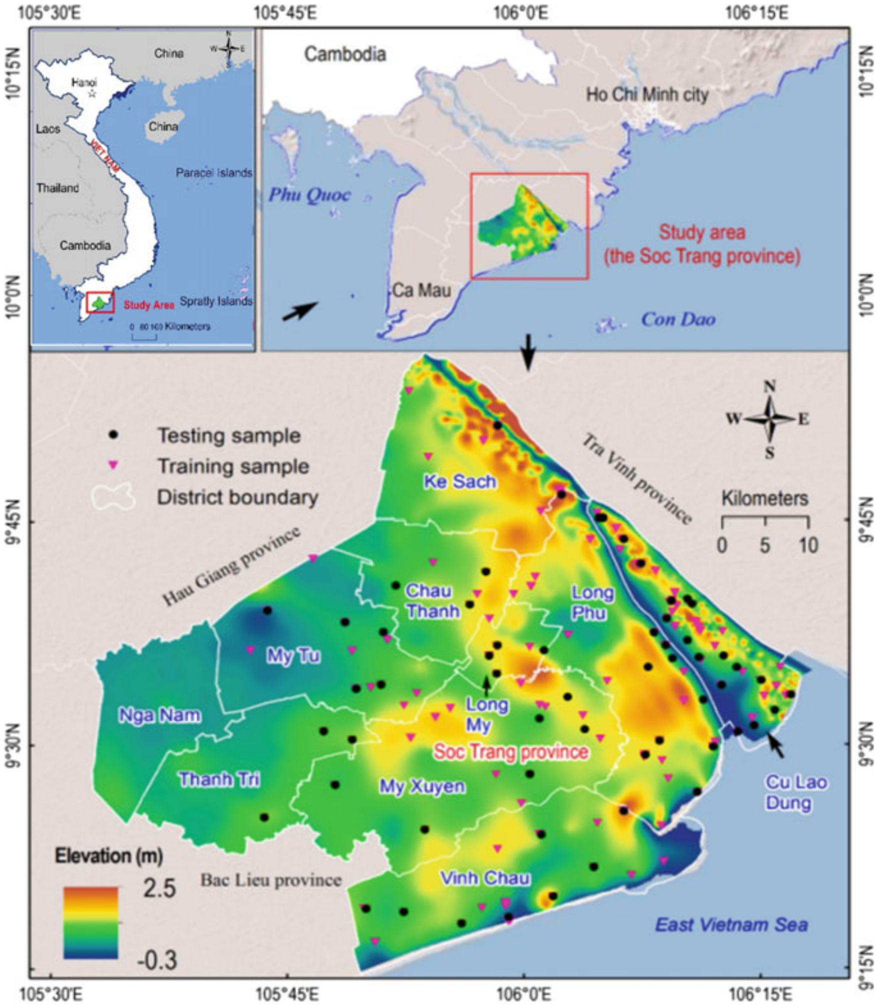

Fig. 25.2 Location of the study area (Soc Trang province), the Vietnamese Mekong River Delta

Soc Trang province has around 1.20 million people in which a majority of the population depends on agriculture for their livelihoods, contributing to 42% of the total GDP in the province (Hoang et al. 2019). Agriculture lands are dominant, accounting for 84.77% (276,690 ha of total area), which includes rice fields (52.98%), fishponds (19.69%), orchards (15.51%), and lands of other vegetable types (6.75%), and other types of land use (Decision No. 108/NQ-CP of the Government 2018).

In the study area, groundwater is used as a dominant source of water for domestic, industrial and agricultural activities, resulting in rapid groundwater level depletion in the irrigated areas (Hoang and Bäumle 2018; Minderhoud et al. 2017). Groundwater salinization has been identified as one of the significant threats to the groundwater resource in this region (An et al. 2018). The extent of groundwater salinization in the study area has recently been increased due to the rapid increase in groundwater demand (Minderhoud et al. 2017; Nam et al. 2019).

The hydrogeological setting of the study area is characterized by a multi-layered aquifer system, formed between the Miocene and Holocene epoch (Wagner et al. 2012; Hung et al. 2019). Groundwater in the Pleistocene aquifers is the primary source of drinking water because these aquifers have high yields and good-quality water compared to other aquifers (An et al. 2018). In this study, we focus on assessing the vulnerability and risk of groundwater in the Pleistocene aquifers to salinity.

25.3.2 Data Preparation and Variables Selection

In this research, 215 groundwater samples from the Pleistocene aquifers were collected between 2013 and 2018 during both the rainy and dry seasons. On-site measurements were conducted to obtain physical parameters such as groundwater temperature T (°C), pH, dissolved oxygen DO, and electrical conductivity EC using the HANNA portable instruments (Hanna Instruments Inc. 2015). The chloride concentration in groundwater samples was analyzed using Ion Liquid Chromatography (Shimadzu Co. Ltd., Japan) at the University of Tsukuba, Japan.

The accumulation of salinity in groundwater is a complex process because it is controlled by influencing factors (Mahlknecht et al. 2017; Kanagaraj et al. 2018). The selection of influencing factors for groundwater salinization prediction based on the possibilities of saltwater migration into aquifers. In the Pleistocene aquifers, groundwater salinity is originated from (1) downward or upward leakage of paleo-saline water (Khaska et al. 2013; Chatton et al. 2016), (2) halite dissolution in the topsoil layer (Walter et al. 2017; Blasco et al. 2019), (3) seawater intrusion (Han and Currell 2018; Kanagaraj et al. 2018; Werner et al. 2013), and (4) irrigation return flow (Essaid and Caldwell 2017; Lapworth et al. 2017; Malki et al. 2017; Tweed et al. 2018). The downward or upward leakages of paleo-saline water may relate to the formation of aquifers, which is further incorporated into the lithology influencing factor. Furthermore, the thicknesses of aquitards, distance to the hydraulic window, distance to fault, fault density, and vertical hydraulic conductivity could also affect the leaking rate (Elmahdy and Mohamed 2013; Liu et al. 2018). Besides, other geographical variables such as distance from main rivers, distance to the drainage and drainage density are also widely considered as influencing factor to groundwater salinity (Winkel et al. 2008). The halite dissolution process is characterized by salt rock/sediment properties, soil type, and horizontal and vertical hydraulic conductivity. Variables which represent the effect human activities on groundwater salinity in the study area are the groundwater level, extraction capacity, well density, extraction density, and operation time. The severity of seawater intrusion may also depend on the distance to the sea, groundwater level, well density, extraction capacity, extraction density, and horizontal hydraulic conductivity (Lee et al. 2016; Yechieli et al. 2019). The four processes mentioned above interact with each other and result in a

Table 25.1 Influencing factors for prediction of groundwater salinity using machine learning models

No.	Explanatory variables	Coding	Unit	Data type
1	Distance to the sea	DTS	km	Numeric
2	Distance to main river	DTR	km	Numeric
3	Distance to drainage	DTD	km	Numeric
4	Drainage density	DD	m/km	Numeric
5	Distance to hydraulic window	DTW	km	Numeric
6	Distance to fault	DTF	km	Numeric
7	Fault density	FD	m/km	Numeric
8	Distance to saline sources	DTS	km	Numeric
9	Temperature of groundwater	T	°C	Numeric
10	Depth of screen well	DSW	m	Numeric
11	Soil properties	AT	#	Ordinal
12	Horizontal hydraulic conductivity	Kh	m/d	Numeric
13	Vertical hydraulic conductivity	Kv	m/d	Numeric
14	Thickness of aquitard	WA	m	Numeric
15	Operation time (of well)	OTW	year	Numeric
16	Well density	WD	well/km^2	Numeric
17	Discharge density	DCD	m^3/km^2	Numeric
18	Extraction capacity	EXC	m^3/d	Numeric
19	Groundwater level	GWL	m.abmsl	Numeric
20	Lithology	LT	#	Ordinal

complex salinization process in the study area (An et al. 2018). Based on the analysis mentioned above, 20 influencing factors were selected for predicting the spatial distribution of salinity in groundwater (Table 25.1).

25.4 The Proposed Methodology for the Prediction of Groundwater Salinity in Coastal Aquifers with Artificial Intelligence Techniques

The modelling framework used in this study is as follows: (1) data pre-processing, (2) feature selection, (3) model parameters, (4) model performance and evaluation, (5) Data post-processing (Fig. 25.3).

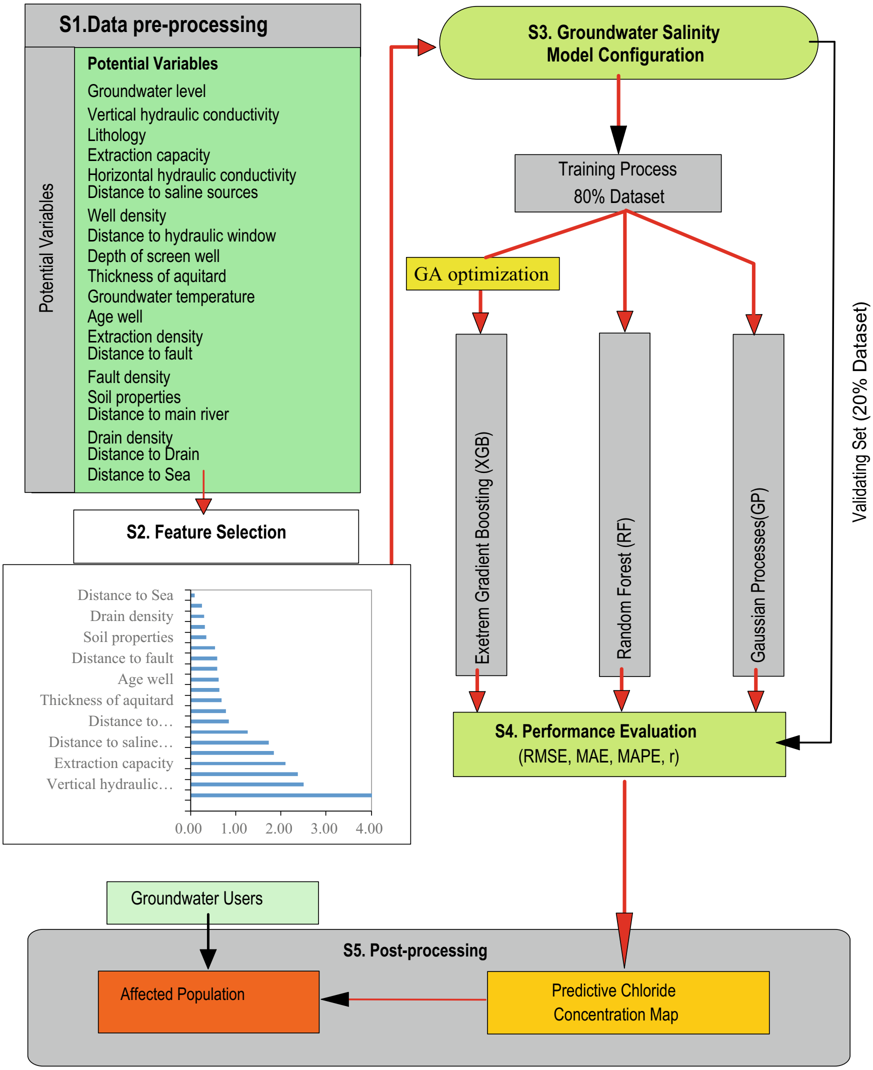

Fig. 25.3 Methodological chart of the present study

25.4.1 Data Pre-processing

Prior to modelling, 215 groundwater samples from middle and lower Pleistocene aquifers were selected, and each sample consists of 20 variables (Table 25.1). The measured Cl^- concentration is assigned as a dependent variable, while the 20 influencing factors are assigned as independent variables. The dataset was then randomly split to training and testing datasets 80% of the dataset was used for training, and 20% of the dataset was used for testing.

Since the influencing factors for predicting groundwater salinization have significantly different ranges, normalization was used to convert the values of numeric columns into a range from 0 to 1 using the following equation:

$$X_n = \frac{X_i - X_{\max}}{X_{\max} - X_{\min}} \tag{25.3}$$

where X_n and X_i represent the moralized and raw training and testing data; X_{max} and X_{min} are the minimum and maximum of the training and testing data.

25.4.2 Feature Selection

As many factors control groundwater salinization processes in coastal aquifers, the selection of influencing factors plays a vital role in reducing time and cost of computation processes and improving the accuracy of prediction results. For several decades, numerous variable selection methods have been applied to identify significant variables before feeding machine learning algorithms to construct predictive models such as filters, wrappers, and embedded techniques (Kohavi and John 1997; Guyon and Elisseeff 2006; Hira and Gillies 2015).

Recently, Random Forests (RF) and its improved algorithms (XGB) have been widely used not only for predicting but also for selecting essential variables as the embedded technique to predictive models (Rodriguez-Galiano et al. 2014; Zeng et al. 2018; Zhao et al. 2019). In this study, the RF algorithm is employed to select input parameters for predicting chloride concentrations in the middle and lower Pleistocene aquifers of the study area. The procedure was followed below steps:

Step 1: Estimation of permutation-based mean squared error (MSE) reduction as Eq. (25.2):

$$MSE^t_{OOB} = \frac{1}{nOOB(t)} \cdot \sum_{i=1}^{nOOB} \left(y_i - \hat{y}_{iOOB,t} \right) \tag{25.4}$$

where MSE_{OOB} is mean squared error, nOOB is the total of out-of-bag (OOB) samples, y_i is the measure Cl^- concentration in groundwater samples, and $\hat{y}_{iOOB,t}$ is the predicted Cl^- concentration of the i-th sample from a decision tree t of OOB samples.

Step 2: Estimation of MSE for permuted input variable x_i using the following equation:

$$MSE^{t}_{OOB}[x_i\,permuted] = \frac{1}{nOOB(t)}.\sum_{i=1}^{nOOB}\left(y_i - \hat{y}_{iOOB,t}\right)[x_i\,permuted] \quad (25.5)$$

Step 3: Estimation of variable importance score for variable x_i using the following equation:

$$VI(x_i) = \frac{1}{T_{tree}}.\sum_{t=1}^{Ttree}\left(MSE^{t}_{OOB}[x_i\,permuted] - MSE^{t}_{OOB}\right) \quad (25.6)$$

25.4.3 Model Configuration and Training

The configuration and training for the three machine learning models are conducted using a training dataset (80% of measured data). For the RF model, the tree-net system is built from 1000 trees with a maximum of 4 nodes per tree and the maximal tree depth of 17. For the GP model, the radial basis function (RBF) kernel and gamma = 0.014 are chosen to predict chloride concentrations in groundwater. In the GO-XGB model, each XGB prediction rule is trained with tenfold cross-validation to identify the number of trees (*ntree*) that minimizes an objective function. The prediction rule is fine-tuned by identifying the optimal combination of hyperparameters that further minimized the objective function for each area. The hyperparameters include the number of base classifiers (*n_estimators*), the maximum depth of each tree (*max_depth*), the learning rate (*eta*), the number of observations in each leaf node of the tree (*min_child_weight*), the minimum loss reduction required to partition further a leaf node on a single tree (*gamma and reg_alpha*), the proportion of observed data were used by XGB algorithm to grow each tree (subsample), and the proportion of predictor variables used at each level of tree splitting (*colsample_bytree*). The *n_estimators* is defined as the number of base classifiers and improper setting of *n_estimators* will result in model failure. The *maximum_tree_depth* was selected appropriately to prevent model complexity. This parameter is crucial in controlling under and over-fitting issues in which too small values of *maximum_tree_depth* will cause underfitting while too large values will result in overfitting. *Learning_rate* represents the weight-reduction factor of each base classifier. *Min_child_weight* represents the weight of the minimum leaf node sample and is used to improve the generalization of the model. The value of gamma ranges from 0 to infinitive, which represents the minimum loss reduction required to make a further partition on a leaf node of the tree. The gamma parameter controls the drop value of the model loss function when the node splits. The subsample controls the proportion of random sampling for each tree, typically between 0.5 and 1. The regularization parameter alpha (*reg_alpha*) denotes the L1 regularization term of the weight, which is used to simplify the complexity of the model.

In this study, the XGB algorithm was used to construct the model and optimizes parameters with GA. The details framework is described in Fig. 25.3. The main parameters of the XGB algorithm that need to be optimized are *max_depth*, *learning_rate*, *min_child_weight*, *subsample*, *alpha*, and *gamma*. After adjusting parameters by a genetic optimization function, we found the best value of these parameters *max_depth* = 15, *learning_rate* = 0.153, *min_child_weight* = 1, *subsample* = 1, *alpha* = 0.005, and *gamma* = 0.0015. In addition, *n_estimators* = s1200, *colsample_bytree* = 0.635, and *n_estimators* = 1200 were selected. The decision rule was retrained and applied to the withheld testing data to predict a new series of count observations and evaluate the accuracy of the decision rule based on the optimal values of the hyperparameters and number of trees. The variable importance of each environmental predictor variable was also obtained using the XGB algorithm.

25.4.4 Performance Assessment

The performance criteria used for evaluating model performance depends on the output variables of each model, e.g., categorical or continuous variable (Tien Bui et al. 2016). For evaluating the model with output values is continuous, performance criteria such as the root mean square error (RMSE), the mean absolute percentage error (MAPE), the mean absolute error (MAE), and Pearson's correlation coefficient (r) (Pham et al. 2019a) are used. Each performance criteria term indicates specific information regarding predictive performance efficiency (Li et al. 2016). RMSE is a quadratic scoring rule that measures the average magnitude of errors. It gives a relatively high weight to large errors; hence, it is most useful when large errors are undesirable. The Mean Absolute Percentage Error (MAPE) is the average of absolute errors divided by actual observation values. MAE measures the average magnitude of errors in a set of predictions without considering their direction. It is a linear score, implying that all individual differences between predictions and corresponding observed values are weighted equally in the average. The r is a measure of the linear correlation between observation and prediction values. RMSE, MAPE, MAE, and rare estimated by the equations (Pham et al. 2019b):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(y_i^{obs} - y_i^{pr}\right)^2}{n}} \tag{25.7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{obs} - y_i^{pr}) \tag{25.8}$$

$$MAPE = \sum_{i=1}^{n}\frac{\left|\frac{y_i^{obs} - y_i^{pr}}{y_i^{obs}}\right|}{n} \times 100 \tag{25.9}$$

$$r = \frac{\sum_{t=1}^{n} (y_i^{obs} - \overline{y_{obs}}) \times (y_i^{pr} - \overline{y_{pr}})}{\sqrt{\sum_{t=1}^{n} (y_i^{obs} - \overline{y_{obs}})} \times \sqrt{\sum_{t=1}^{n} (y_i^{pr} - \overline{y_{pr}})}} \tag{25.10}$$

where y_i^{obs} and y_i^{pr} are measured and predicted Cl^- concentration in observation i, and n is the number of observations. Higher values of rare preferred, i.e. close to 1, means better model performance and regression line fits the data well. Conversely, the lower values of RMSE, MAPE, and MAE values the better model performances.

25.4.5 Generating Groundwater Salinity Map

The results from the three machine learning models are then used to create chloride concentration maps. Prediction maps are constructed with four main steps as follows: (i) interpolating chloride concentrations in groundwater based on prediction results, (ii) reclassifying chloride concentrations based on the drinking water standard from WHO, (iii) estimating the salinity affected area, and (iv) estimating the number of people in each class of salinity affected area. In the first step, the predicted chloride concentrations are interpolated to create maps using the Kriging method by Spatial Analysis Tool in ArcGIS 10.3. In the second step, the interpolated results are reclassified into four main classes, including low ($Cl^- < 250$ mg/L), moderate ($250 \leq Cl^- \leq 500$ mg/L), high ($500 \leq Cl^- \leq 1000$ mg/L), and high ($Cl^- > 1000$ mg/L). In the third step, the salinity affected area for each class of the salinity concentration in groundwater was calculated using geometry functions in ArcGIS 10.3. In the final step, the numbers of people within each salinity affected area was estimated based on the salinity-affected areas and population density.

25.5 Result and Discussion

25.5.1 Feature Selection for the Groundwater Salinity Modelling

The results in Table 25.2 showed the variable importance selection with the permutation based MSE decreased values ranged from 4.03 to 0.69.

In the study area, the top ten most important influencing factors are groundwater level (4.03), vertical hydraulic conductivity (2.50), lithology (2.37), extraction capacity (2.10), horizontal hydraulic conductivity (1.85), distance to saline sources (1.73), well density (1.26), distance to hydraulic windows (0.85), depth of screen wells (0.79), and thickness of aquitards (0.69). The result reveals that groundwater salinization depends not only on hydrogeological features (vertical and horizontal hydraulic conductivities, lithology, paleo-saline sources, hydraulic connection, depth

Table 25.2 Variable importance (permutation based MSE decreased)

No.	Variable	Permutation-based MSE decreased	Number of nodes used	Ranking
1	Groundwater level	4.03	636	1
2	Vertical hydraulic conductivity	2.50	35	2
3	Lithology	2.37	50	3
4	Extraction capacity	2.10	281	4
5	Horizontal hydraulic conductivity	1.85	260	5
6	Distance to saline sources	1.73	379	6
7	Well density	1.26	242	7
8	Distance to hydraulic window	0.85	529	8
9	Depth of screen well	0.79	518	9
10	Thickness of aquitard	0.69	376	10
11	Groundwater temperature	0.64	354	11
12	Age well	0.62	135	12
13	Extraction density	0.59	208	13
14	Distance to fault	0.59	400	14
15	Fault density	0.54	158	15
16	Soil properties	0.35	83	16
17	Distance to main river	0.32	863	17
18	Drain density	0.30	686	18
19	Distance to Drain	0.25	628	19
20	Distance to Sea	0.10	2199	20

of screen well, and thickness of aquitard) but also groundwater extraction practices (groundwater level, extraction capacity, well density). These influencing factors also play an important role in transportation processes of other solutes such as arsenic, fluoride and nitrate in groundwater (Ransom et al. 2017; Winkel et al. 2008; Podgorski et al. 2018). The hydrogeological features influence on moving of saline groundwater from shallow to deeper aquifers (Hung et al. 2019) while groundwater exploitation activities exacerbate groundwater salinization (Hoang and Bäumle 2018; An et al. 2018). The result may also suggest that saline groundwater leaking from upper layers to lower layers is a dominant process, resulting in an increase of chloride concentration in groundwater of the study area. Hydraulically, an increase hydraulic gradient due to groundwater depletion coupled with high vertical hydraulic conductivity, think aquitard, and high-density gradients cause an increase of vertical flow rate as shown in the following equations (Ma et al. 2015).

$$q_v = -\delta \times K_v \left[\frac{h_{up} - h_{low}}{\Delta L} + \varepsilon \left(\frac{C_{up} + C_{low}}{2} \right) \right] \quad (25.11)$$

$$\delta = \frac{\mu_0}{\mu} = 1 - \xi \times \varepsilon \quad (25.12)$$

where: δ—the ratio of the dynamic viscosity of freshwater to seawater; K_v is a vertical hydraulic conductivity (m d^{-1}); h_{up} and h_{low} denote the freshwater equivalent hydraulic heads at upper and lower layers (m), ΔL is the distance from upper to lower layers (m); μ_0 and μ denote the dynamic viscosity (kg m^{-1} d^{-1}); ξ is a constant; C_{up} is average observed salinity of pore water in upper aquifers (kg/m^3); C_{low} is observed salinity of pore water in lower aquifers (kg/m^3), and ε is a constant. The similar findings were also observed in other coastal aquifers in the world (Chatton et al. 2016; Cary et al. 2015; Delsman et al. 2014; Larsen et al. 2017), which indicated strong influences of over groundwater exploitation on seawater intrusion in coastal aquifers (Yechieli et al. 2019; Yu and Michael 2019; Han et al. 2015).

The other major influencing factors have permutation based MSE values from 0.64 for groundwater temperature to 0.10 for distance to the sea. It was noted that the distance to the sea had a little score value of 0.10, indicating less contribution to groundwater salinization processes. This result may suggest that direct seawater intrusion from the sea to coastal aquifers of the study is not dominant in the study area.

25.5.2 Model Performance Evaluation and Comparison

In this study, the predictive models for groundwater salinization are built using the training and the testing datasets, drawing upon a total of 215 observation wells and 20 variables. The results of the goodness-of-fit assessment of the three machine learning algorithms-based models including the GO-XGB model, RF model and the GP model for both training and testing steps are shown in Fig. 25.4 and summarized in Tables 25.3 and 25.4, and respectively.

The training model performance (Table 25.3) shows that the GO-XGB model has the lowest value RMSE = 141.042 mg/L, followed by the RF (RMSE = 176.179 mg/L) and GP (RMSE = 176.179 mg/L) models. The similar trend is also observed in MAE and MAPE for the GO-XGB (MAE = 4.864, MAPE = 2.070), RF (MAE = 58.286 mg/L, MAPE = 29.410 mg/L) and GP (MAE = 71.802 mg/L, MAPE = 61.42 mg/L). In contrast, the GO-XGB model has the highest r-value of 0.999 compared to that of RF (r = 0.786) and Gaussian Processes (r = 0.882).

In the testing step, the results of the predictive models are validated by using the testing dataset consisted of 20% random samples from the original dataset (Fig. 25.4). The testing results show that the GO-XGB model has the highest performance compared to the RF and GP models (Table 25.4). For example, GO-XGB has the best result of r = 0.787, followed by the RF model (r = 0.596) and the GP model

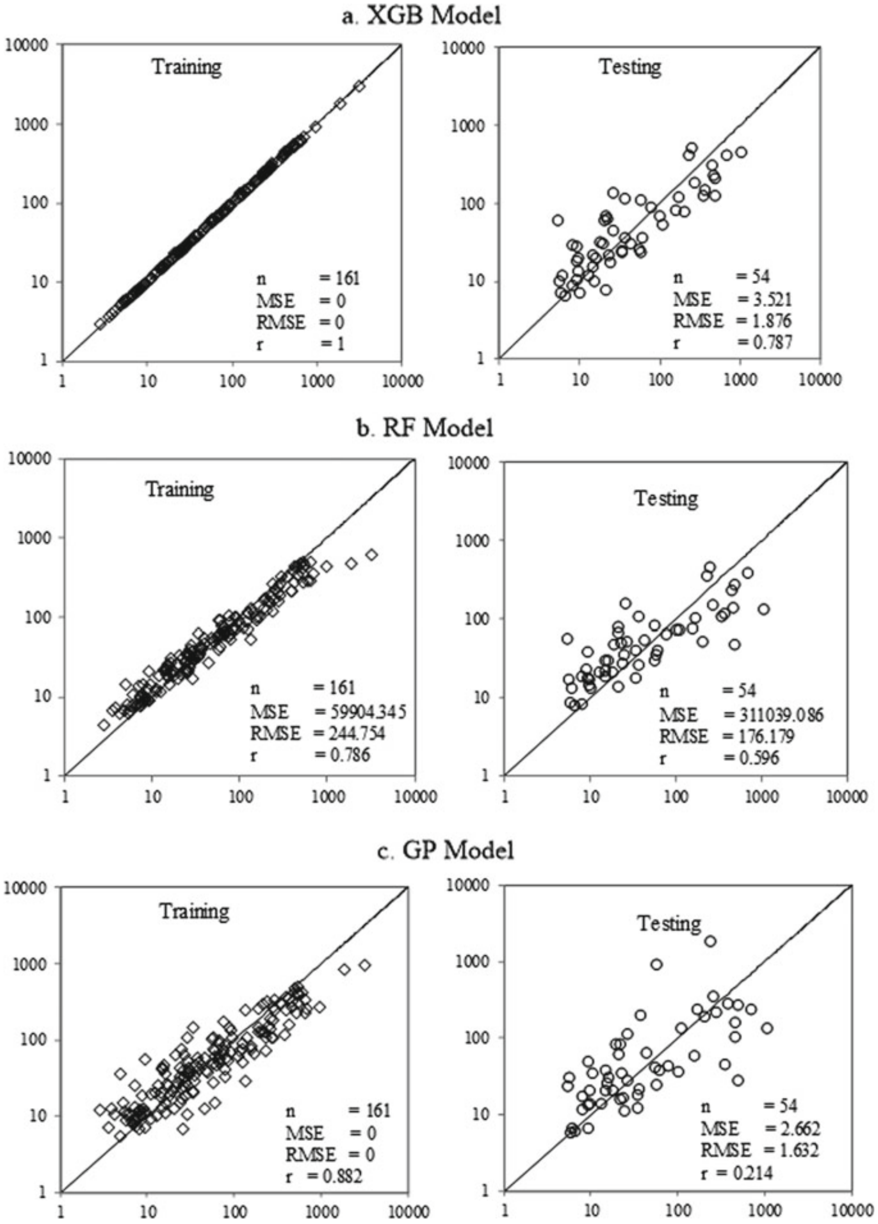

Fig. 25.4 Observed versus predicted chloride concentration for training and test data for **a** GO-XGB, **b** RF, and GP model

Table 25.3 Goodness-of-fit of the ground water salinity models on the training dataset

Statistical metrics	GO-XGB	RF	Gaussian processes
RMSE	18.450	244.754	219.329
MAE	4.864	58.286	71.802
MAPE	2.070	29.410	61.42
r	0.999	0.786	0.882

Table 25.4 Prediction performance of the ground water salinity models using the validation dataset

Statistical metrics	GO-XGB	RF	Gaussian processes
RMSE	141.042	176.179	305.782
MAE	74.993	84.708	127.355
MAPE	87.250	95.780	130.840
r	0.787	0.596	0.214

(r = 0.214). Similarly, the GO-XGB model shows the lowest values of RMSE = 141.042 mg/L, MAE = 74.993 mg/L, and MAPE = 87.250 mg/L, followed by the RF (RMSE = 176.179 mg/L, MAE = 84.708 mg/L, MAPE = 95,780 mg/L) and GP (RMSE = 305.782 mg/L, MAE = 127.355 mg/L, MAPE = 130.840 mg/L) models.

Overall, the GO-XGB model produces an excellent predictive performance with the highest value of r = 0.99 and r = 0.787 for training and validation steps among three predictive models. Likewise, this model also has the lowest values of RMSE, MAE, and MAPE compared to the RF and GP models in both training and validation steps.

Although we have considered various influencing factors to provide the accurate prediction of groundwater salinity in a coastal area of the Mekong River Delta, however, the processes of seawater intrusion into fresh aquifers depend not only human activities but also natural variations. Therefore, for broader applicability, these models would be required to include additional influencing factors such as the regional groundwater flow system, tidal fluctuation, climate change, and sea-level rise. Also, the performance of prediction models may have to compare with numerical models and other stochastic models.

25.5.3 Mapping Salt-Groundwater-Affected Area

In general, the average results obtained from the three machine learning models, including the GO-XGB (Fig. 25.5), RF (Fig. 25.6), and the GP models (Fig. 25.7), shows the main salinity-affected region, extending from the My Thanh River to the Central of Soc Trang City. It was noted that the prediction results from GO-XGB model strongly agree with salinity observation in this study (Fig. 25.8) and previous studies (An et al. 2018). Accordingly, high chloride concentrations which exceed the limited standard for drinking water $Cl^- > 250$ mg/L is predicted in the areas with to paleo-saline sources, high extraction rates, and significant groundwater level depletion. The severely affected areas are the Tran De estuary, the My Thanh river and the central region including Soc Trang city and My Xuyen district where chloride concentrations in wells elevate to 2000 mg/L. Surprisingly, low chloride concentrations ($Cl^- < 250$ mg/L) in groundwater is predicted in coastal areas even if in the production wells located just around 2 km from the sea and at −10.5 m below the mean sea level (m.a.m.sl). Meanwhile, Soc Trang city, which locates far from the sea approximately 40 km, is predicted to have high chloride concentrations in groundwater. This reveals that processes of salinity accumulation in aquifers are very complex, depending not only on natural processes but also human-induced activities.

The spatial distributions of affected areas with moderate and high chloride concentration are relative differences among models. For example, in the GO-XGB model, the affected area is predicted to extend from the coastal line to the central area of the study (Fig. 25.6).

In addition, the profoundly affected area is observed in the substantial groundwater extraction locations. These locations are located close to the paleo-saline groundwater sources coupled, and these areas also have high groundwater extraction rates and significant groundwater level depletion. This indicates that these influencing factors play an essential role in increasing chloride concentrations in groundwater. The similar finding is also in-line with recent studies (Hoang and Bäumle 2019; Tran

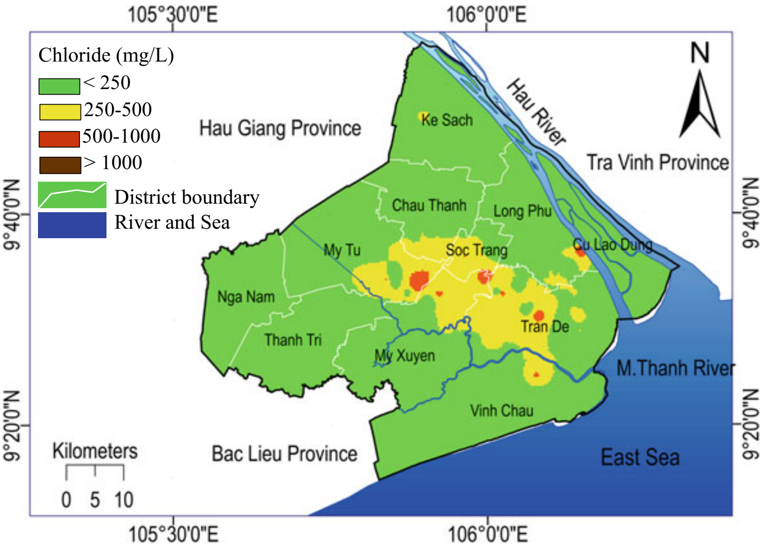

Fig. 25.5 Predicted chloride concentration in groundwater of the study area using GO-XGB model

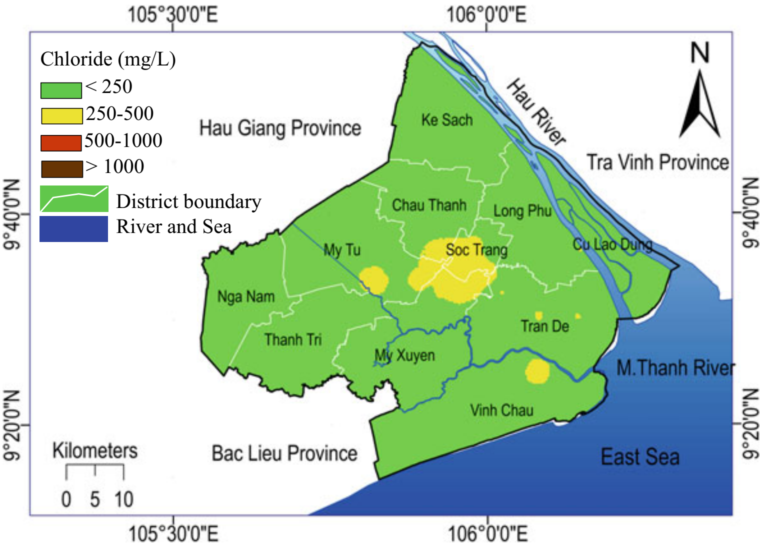

Fig. 25.6 Predicted chloride concentration in groundwater of the study area using RF model

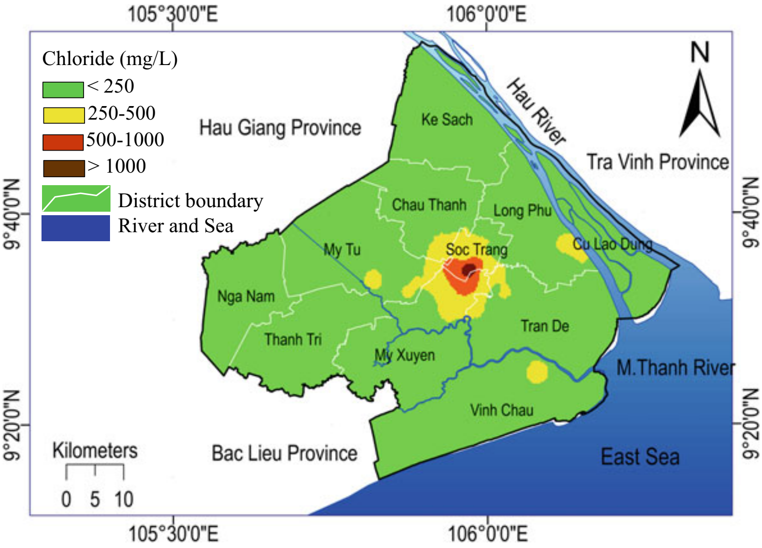

Fig. 25.7 Predicted chloride concentration in groundwater of the study area using GP model

et al. 2019). Conversely, the results from the RF (Fig. 25.6) and the GP models (Fig. 25.7) show that the moderately affected areas are the central area of the study region.

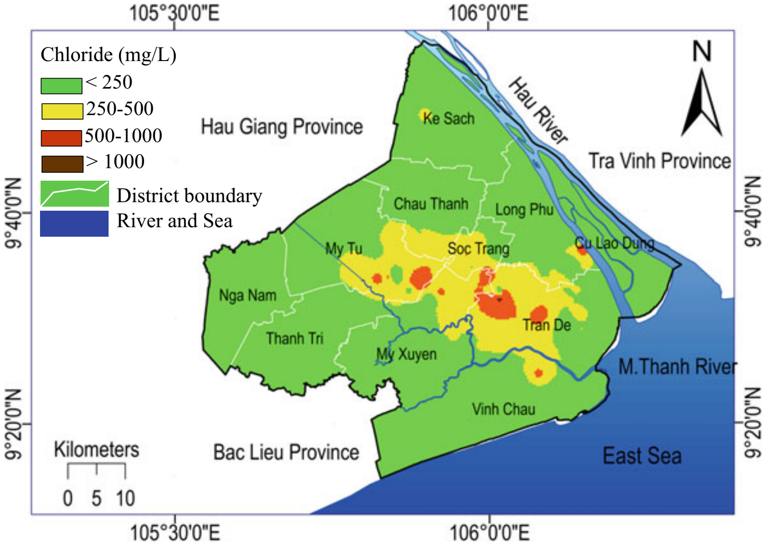

Fig. 25.8 Measured chloride concentration in groundwater of the study area

The three models provide different predictions in the affected area (Table 25.5). The RF model predicted the largest affected area (3118.50 km^2) followed by the GP model (3055.35 km^2) and GO-XGB (2879.0 km^2) with low chloride concentration (Cl$^-$ < 250 mg/L). Meanwhile, the largest affected-areas with moderate-high chloride concentration (Cl$^-$ = 250–500 mg/L) are observed by the GO-XGB model (433 km^2), the GP model (256.65 km^2), and the RF model (193.50 km^2). Both the GO-XGB model and the GP models predicted the large affected-areas with high (Cl = 500–1000 mg/L) and very chloride concentration (>1000 mg/L) while RF model predicted non-affected areas of high and very high chloride concentration.

25.6 Concluding Remarks

In this study, three advanced machine learning models, including GO-XG, RF, and GP, were employed to predict chloride concentration in groundwater and assess impacts of salinity on water users in a coastal area of the Mekong River Delta, Vietnam. Twenty influencing factors were evaluated using the RF model based on score estimation. The most influenced factors to high salinity are related to both groundwater exploitation (groundwater level depletion, extraction capacity, and well density) and hydrogeological features (vertical hydraulic conductivity lithology, horizontal hydraulic conductivity, distance to the saline source, distance to the hydraulic window, depth of screen well, and thickness of aquitard). This finding confirms

Table 25.5 Predictive results of affected areas (in km^2) following four classes of chloride concentration in groundwater

Statistical metrics	Low (Cl < 250 mg/L)	Moderate (250 ≤ Cl$^-$ ≤ 500 mg/L)	High (500 < Cl$^-$ ≤ 1000 mg/L)	Very high (Cl$^-$ > 100 mg/L)	% Affected area
GO-XGB	2879.00	433.00	42.40	0.88	14.0
RF	3118.50	193.50	0	0	6.0
Gaussian processes	3055.35	256.65	34.46	4.86	9.0
Average	3017.62	294.38	25.62	1.91	10.0

previous studies in which groundwater exploitation is one of the most important influencing factors to seawater intrusion in coastal lowland regions.

All three models perform well in predicting the probability of groundwater salinity. However, the GO-XGB model provides the highest accuracy prediction with RMSE = 18.450, MAE = 4.864, MAPE = 2.070, and r = 0.999 compared to the GP model (RMSE = 219.329, MAE = 71.329, MAPE = 61.42, and r = 0.882) and the RF model (RMSE = 244.754, MAE = 58.286, MAPE = 29.410, and r = 0.786). This indicated that GO-XGB model could be a useful tool to predict groundwater salinization in the coastal aquifers.

All three models predicted that approximately 35% of the total population might have to use groundwater with chloride concentration exceeding the WHO drinking water standard ($Cl^- > 250$ mg/L). More seriously, urban areas are close to paleo-saline sources. While the thicknesses of aquitards are thin and groundwater levels deplete quickly, leaking paleo-saline becomes more server and cause groundwater salinization. This is stimulated by the hydraulic connection between aquifers and over groundwater exploitation. Given the rapid increase of water demand, significant groundwater depletion and unpredictable impacts of climate change and sea-level rise, immediate actions must be taken by the water authorities to find a suitable solution to this environmental crisis.

Conflicts of Interest The authors declare no conflict of interest.

References

Abdelhamid H et al (2016) Simulation of seawater intrusion in the Nile Delta aquifer under the conditions of climate change, vol 47

An TD et al (2018) Isotopic and hydrogeochemical signatures in evaluating groundwater quality in the Coastal Area of the Mekong Delta, Vietnam. In: Bui DT et al (eds) Advances and applications in geospatial technology and earth resources: proceedings of the international conference on geo-spatial technologies and earth resources 2017. Springer International Publishing, Cham, pp 293–314

Azimi S, Moghaddam MA, Hashemi Monfared SA (2018) Large-scale association analysis of climate drought and decline in groundwater quantity using Gaussian process classification (case study: 609 study area of Iran). J Environ Health Sci Eng 16(2):129–145

Banerjee P et al (2011) Artificial neural network model as a potential alternative for groundwater salinity forecasting. J Hydrol 398(3–4):212–220

Behera AK et al (2019) Identification of seawater intrusion signatures through geochemical evolution of groundwater: a case study based on coastal region of the Mahanadi delta, Bay of Bengal, India. Nat Hazards 97(3):1209–1230

Blasco M, Auqué LF, Gimeno MJ (2019) Geochemical evolution of thermal waters in carbonate—evaporitic systems: the triggering effect of halite dissolution in the dedolomitisation and albitisation processes. J Hydrol 570:623–636

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brouwer R et al (2018) Economic valuation of groundwater protection using a groundwater quality ladder based on chemical threshold levels. Ecol Ind 88:292–304

Carretero S et al (2013) Impact of sea-level rise on saltwater intrusion length into the coastal aquifer, Partido de La Costa, Argentina. Cont Shelf Res 61–62:62–70

Cary L et al (2015) Origins and processes of groundwater salinization in the urban coastal aquifers of Recife (Pernambuco, Brazil): a multi-isotope approach. Sci Total Environ 530–531:411–429

Chatton E et al (2016) Glacial recharge, salinisation and anthropogenic contamination in the coastal aquifers of Recife (Brazil). Sci Total Environ 569–570:1114–1125

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, San Francisco, California, USA, pp 785–794

Chen W et al (2019) Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling landslide susceptibility. CATENA 172:212–231

Criminisi A (2011) Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends Comput Graph Vis 7(2–3):81–227

de Graaf IEM et al (2019) Environmental flow limits to global groundwater pumping. Nature 574(7776):90–94

Delsman JR et al (2014) Paleo-modeling of coastal saltwater intrusion during the Holocene: an application to the Netherlands. Hydrol Earth Syst Sci 18(10):3891–3905

Elmahdy SI, Mohamed MM (2013) Influence of geological structures on groundwater accumulation and groundwater salinity in Musandam Peninsula, UAE and Oman. Geocarto Int 28(5):453–472

Essaid HI, Caldwell RR (2017) Evaluating the impact of irrigation on surface water—groundwater interaction and stream temperature in an agricultural watershed. Sci Total Environ 599–600:581–596

Famiglietti JS (2014) The global groundwater crisis. Nat Clim Chang 4(11):945–948

Ferguson G, Gleeson T (2012) Vulnerability of coastal aquifers to groundwater use and climate change. Nat Clim Chang 2(5):342–345

Forrest S (1993) Genetic algorithms: principles of natural selection applied to computation. Science 261(5123):872–878

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

Gallardo AH, Marui A (2007) Modeling the dynamics of the freshwater-saltwater interface in response to construction activities at a coastal site. Int J Environ Sci Technol 4(3):285–294

Giambastiani BMS et al (2018) Forest fire effects on groundwater in a coastal aquifer (Ravenna, Italy). Hydrol Process 32(15):2377–2389

Guhl F et al (2006) Geometry and dynamics of the freshwater—seawater interface in a coastal aquifer in southeastern Spain. Hydrol Sci J 51(3):543–555

Guyon I, Elisseeff A (2006) An introduction to feature extraction, in feature extraction. Springer, Berlin, Heidelberg, pp 1–25

Hall J, Rasmussen C, Maciejowski J (2012) Modelling and control of nonlinear systems using Gaussian processes with partial model information. In: 2012 IEEE 51st IEEE conference on decision and control (CDC)

Han D, Currell MJ (2018) Delineating multiple salinization processes in a coastal plain aquifer, northern China: hydrochemical and isotopic evidence. Hydrol Earth Syst Sci 22(6):3473–3491

Han D, Post VEA, Song X (2015) Groundwater salinization processes and reversibility of seawater intrusion in coastal carbonate aquifers. J Hydrol 531:1067–1080

Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 2015:1–13

Hoa PV et al (2019) Soil salinity mapping using SAR Sentinel-1 data and advanced machine learning algorithms: a case study at Ben Tre Province of the Mekong River Delta (Vietnam). Rem Sens 11(2):128

Hoang HT, Bäumle R (2018) Complex hydrochemical characteristics of the middle–upper Pleistocene aquifer in Soc Trang Province, Southern Vietnam. Environ Geochem Health

Hoang HT, Bäumle R (2019) Complex hydrochemical characteristics of the middle-upper Pleistocene aquifer in Soc Trang Province, Southern Vietnam. Environ Geochem Health 41(1):325–341

Hoang LP et al (2019) The Mekong's future flows under multiple drivers: how climate change, hydropower developments and irrigation expansions drive hydrological changes. Sci Total Environ 649:601–609

Isazadeh M, Biazar SM, Ashrafzadeh A (2017) Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. Environ Ear Sci 76(17):610–614

Javadi A et al (2015) Multi-objective optimization of different management scenarios to control seawater intrusion in coastal aquifers. Water Resour Manage 29(6):1843–1857

Jennings PC et al (2019) Genetic algorithms for computational materials discovery accelerated by machine learning. NPJ Comput Mater 5(1):46–52

Johnson NE, Bonczak B, Kontokosta CE (2018) Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. Atmos Environ 184:9–16

Kagabu M et al (2020) Describing coseismic groundwater level rise using tank model in volcanic aquifers, Kumamoto, southern Japan. J Hydrol 582:124464-14

Kanagaraj G et al (2018) Hydrogeochemical processes and influence of seawater intrusion in coastal aquifers south of Chennai, Tamil Nadu, India. Environ Sci Pollut Res 25(9):8989–9011

Kaur L et al (2020) Groundwater potential assessment of an alluvial aquifer in Yamuna sub-basin (Panipat region) using remote sensing and GIS techniques in conjunction with analytical hierarchy process (AHP) and catastrophe theory (CT). Ecol Ind 110:105850-19

Khaska M et al (2013) Origin of groundwater salinity (current seawater vs. saline deep water) in a coastal karst aquifer based on Sr and Cl isotopes. Case study of the La Clape massif (southern France). Appl Geochem 37:212–227

Kim IH, Yang J-S (2018) Prioritizing countermeasures for reducing seawater-intrusion area by considering regional characteristics using SEAWAT and a multicriteria decision-making method. Hydrol Process 32(25):3741–3757

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1):273–324

Kopsiaftis G et al (2019) Gaussian process regression tuned by Bayesian optimization for seawater intrusion prediction. Comput Intell Neurosci 2019:2859429-12

Lal A, Datta B (2019) Multi-objective groundwater management strategy under uncertainties for sustainable control of saltwater intrusion: Solution for an island country in the South Pacific. J Environ Manage 234:115–130

Lapworth DJ et al (2017) Groundwater quality in the alluvial aquifer system of northwest India: new evidence of the extent of anthropogenic and geogenic contamination. Sci Total Environ 599–600:1433–1444

Larsen F et al (2017) Groundwater salinity influenced by Holocene seawater trapped in incised valleys in the Red River delta plain. Nat Geosci 10(5):376–381

Lee S, Currell M, Cendón DI (2016) Marine water from mid-Holocene sea level highstand trapped in a coastal aquifer: evidence from groundwater isotopes, and environmental significance. Sci Total Environ 544:995–1007

Li Y et al (2016) A fully coupled depth-integrated model for surface water and groundwater flows. J Hydrol 542:172–184

Lim S, Chi S (2019) Xgboost application on bridge management systems for proactive damage estimation. Adv Eng Inform 41:100922-14

Liu Y et al (2018) Geographically weighted temporally correlated logistic regression model. Sci Rep 8(1):1417-14

Ma Q et al (2015) Estimation of seawater–groundwater exchange rate: case study in a tidal flat with a large-scale seepage face (Laizhou Bay, China). Hydrogeol J 23(2):265–275

Ma Y et al (2019a) Characteristics of groundwater pollution in a vegetable cultivation area of typical facility agriculture in a developed city. Ecol Ind 105:709–716

Ma X, Xu F, Chen B (2019b) Interpolation of wind pressures using Gaussian process regression. J Wind Eng Ind Aerodyn 188:30–42

Mahlknecht J et al (2017) Assessing seawater intrusion in an arid coastal aquifer under high anthropogenic influence using major constituents, Sr and B isotopes in groundwater. Sci Total Environ 587–588:282–295

Mahmoodzadeh D, Karamouz M (2019) Seawater intrusion in heterogeneous coastal aquifers under flooding events. J Hydrol 568:1118–1130

Malki M et al (2017) Impact of agricultural practices on groundwater quality in intensive irrigated area of Chtouka-Massa, Morocco. Sci Total Environ 574:760–770

Melloul AJ, Goldenberg LC (1997) Monitoring of seawater intrusion in coastal aquifers: basics and local concerns. J Environ Manage 51(1):73–86

Minderhoud PSJ et al (2017) Impacts of 25 years of groundwater extraction on subsidence in the Mekong delta, Vietnam. Environ Res Lett 12(6):064006-13

Mohanty AK, Rao VVSG (2019) Hydrogeochemical, seawater intrusion and oxygen isotope studies on a coastal region in the Puri District of Odisha, India. CATENA 172:558–571

Nadiri AA et al (2018) Mapping specific vulnerability of multiple confined and unconfined aquifers by using artificial intelligence to learn from multiple DRASTIC frameworks. J Environ Manage 227:415–428

Naghibi SA, Pourghasemi HR, Dixon B (2015) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess 188(1):44–71

Nam NDG et al (2019) Assessment of groundwater quality and its suitability for domestic and irrigation use in the coastal zone of the Mekong Delta, Vietnam. In: Stewart MA, Coclanis PA (eds) Water and power: environmental governance and strategies for sustainability in the lower Mekong Basin. Springer International Publishing, Cham, pp 173–185

Nishanthiny SC et al (2010) Irrigation water quality based on hydro chemical analysis, Jaffna, Sri Lanka. Am Eurasian J Agric Environ Sci 7(1):100–102

Paine JG (2003) Determining salinization extent, identifying salinity sources, and estimating chloride mass using surface, borehole, and airborne electromagnetic induction methods. Water Resour Res 39(3):3–10

Park J, Kwock CK (2015) Sodium intake and prevalence of hypertension, coronary heart disease, and stroke in Korean adults. J Ethnic Foods 2(3):92–96

Park J, Sandberg IW (1991) Universal approximation using radial-basis-function networks. Neural Comput 3(2):246–257

Pham BT et al (2019a) Hybrid computational intelligence models for groundwater potential mapping. CATENA 182:104101–104113

Pham BT et al (2019b) A novel artificial intelligence approach based on multi-layer perceptron neural network and biogeography-based optimization for predicting coefficient of consolidation of soil. CATENA 173:302–311

Podgorski JE et al (2018) Prediction modeling and mapping of groundwater fluoride contamination throughout India. Environ Sci Technol 52(17):9889–9898

Ransom KM et al (2017) A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Sci Total Environ 601–602:1160–1172

Rasmussen CE (2003) Gaussian processes in machine learning. In: Bousquet O, von Luxburg U, Rätsch G (eds) Advanced lectures on machine learning: ML summer schools 2003, Canberra, Australia, 2–14, 2003, Tübingen, Germany, 4–16 Aug 2003. Springer, Berlin, Heidelberg, pp 63–71

Rizeei HM et al (2019) Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. J Hydrol 579:124172-11

Rodriguez-Galiano V et al (2014) Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci Total Environ 476–477:189–206

Roy Dilip K, Datta B (2017) Multivariate adaptive regression spline ensembles for management of multilayered coastal aquifers. J Hydrol Eng 22(9):04017031-13

Sajedi-Hosseini F et al (2018) A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci Total Environ 644:954–962

Scholkopf B et al (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Signal Process 45(11):2758–2765

Sreekanth J, Datta B (2010) Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models. J Hydrol 393(3–4):245–256

Stein S et al (2019) The effect of pumping saline groundwater for desalination on the fresh–saline water interface dynamics. Water Res 156:46–57

Sun Y et al (2016) Technical note: Application of artificial neural networks in groundwater table forecasting—a case study in a Singapore swamp forest. Hydrol Earth Syst Sci 20(4):1405–1412

Tien Bui D et al (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13(2):361–378

Tran DA et al (2019) Stable isotope characteristics of water resources in the coastal area of the Vietnamese Mekong Delta. Isot Environ Health Stud 55(6):566–587

Tweed S et al (2018) Impact of irrigated agriculture on groundwater resources in a temperate humid region. Sci Total Environ 613–614:1302–1316

Van Hung P et al (2019) Paleo-hydrogeological reconstruction of the fresh-saline groundwater distribution in the Vietnamese Mekong Delta since the late Pleistocene. J Hydrol Reg Stud 23:100594-22

Vineis P, Chan Q, Khan A (2011) Climate change impacts on water salinity and health. J Epidemiol Glob Health 1(1):5–10

Voss CI, Souza WR (1987) Variable density flow and solute transport simulation of regional aquifers containing a narrow freshwater-saltwater transition zone. Water Resour Res 23(10):1851–1866

Wagner F, Tran VB, Renaud FG (2012) Groundwater resources in the Mekong Delta: availability, utilization and risks. In: Renaud FG, Kuenzer C (eds) The Mekong delta system: interdisciplinary analyses of a River Delta. Springer, Netherlands, Dordrecht, pp 201–220

Walter J et al (2017) The influence of water/rock—water/clay interactions and mixing in the salinization processes of groundwater. J Hydrol Reg Stud 13:168–188

Werner AD et al (2013) Seawater intrusion processes, investigation and management: recent advances and future challenges. Adv Water Resour 51:3–26

Winkel L et al (2008) Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. Nat Geosci 1(8):536–542

Yadav B et al (2018) Data-based modelling approach for variable density flow and solute transport simulation in a coastal aquifer. Hydrol Sci J 63(2):210–226

Yechieli Y et al (2019) Recent seawater intrusion into deep aquifer determined by the radioactive noble-gas isotopes 81Kr and 39Ar. Earth Planet Sci Lett 507:21–29

Yu X, Michael HA (2019) Mechanisms, configuration typology, and vulnerability of pumping-induced seawater intrusion in heterogeneous aquifers. Adv Water Resour 128:117–128

Zeng X et al (2018) Identifying key factors of the seawater intrusion model of Dagu river basin, Jiaozhou Bay. Environ Res 165:425–430

Zhao X et al (2019) Identifying N6-methyladenosine sites using extreme gradient boosting system optimized by particle swarm optimizer. J Theor Biol 467:39–47