Contents lists available at ScienceDirect

Journal of Hydro-Environment Research

journal homepage: http://ees.elsevier.com

Development of an efficient conjunctive meta-model-based decision-making framework for saltwater intrusion management in coastal aquifers

Ali Ranjbar [a], Najmeh Mahjouri [a,*], Claudia Cherubini [b]

[a] Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran
[b] Department of Physics and Earth Sciences, University of Ferrara, Ferrara, Italy

ARTICLE INFO

Keywords
Simulation-optimization model
K-means clustering
Surrogate model
Random subspace model (RSM)
Nash bargaining theory
SEAWAT
M5 tree

ABSTRACT

This paper presents an integrated framework for management of aquifers threatened by saltwater intrusion (SI). In this framework, SEAWAT model is used for simulating the density dependent groundwater flow. Three meta-models based on the artificial neural network (ANN), M5 tree and random subspaces model (RSM) are developed, as surrogate models for SEAWAT to accurately simulate the groundwater response to different pumping and recharge scenarios. Various patterns of recharge to and discharge from aquifer are used to generate a database for training the mentioned surrogate models. To decrease the number of training parameters, the aquifer area is divided into different zones using k-means clustering technique (KMC). Additionally, a conjunctive model (CM) using a combination of the three surrogate models is proposed to enhance the accuracy of the simulation. It is then integrated with the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) with the objectives of maximizing pumping rates and minimizing SI length. Next, the socially optimal scenarios are selected from the obtained Pareto-front using the Nash bargaining theory. The performance of the proposed model is evaluated by applying it to the Kahak aquifer, Iran, which is subjected to SI. The results show that the conjunctive model using KMC technique predicts SI length with a comparable accuracy and results in 95% reduction in runtime compared to a simulation-optimization (SO) model.

1. Introduction

Optimal management of coastal aquifers is challenging due to the non-linear nature of SI, which is considered as a major threat for groundwater quality. Most management models include a SI simulation model. To simulate the temporal and spatial distribution of groundwater head and salinity in the aquifer, the numerical or analytical simulation models are coupled with optimization algorithms (Bhattacharjya and Datta, 2005; Banerjee et al. 2011) known as simulation–optimization (SO) model. The objective of the SO models mostly include maximizing profit and minimizing cost of pumping operation and these models usually have many decision variables (Sreekanth and Datta, 2015).

The SI process can be simulated using analytical approaches based on sharp interface assumption or variable density flow theory considering a transition zone between saltwater and freshwater (Dausman et al. 2010). The simulation in variable density models is mostly done using numerical methods and based on field salinity data (Werner et al. 2013). Extensive studies have been implemented on management of aquifers subjected to SI using integration of density-dependent numerical codes with optimization algorithms (Mantoglou and Papantoniou, 2008; Dhar and Datta, 2009; Gaur et al. 2011). However, the iterative process of simulation–optimization to converge to optimum solutions is time-consuming (Masoumi and Kerachian, 2008, Mahjouri and Kerachian, 2011, Christelis and Mantoglou, 2016a,b). Surrogate models are efficient tools to decrease the computational time needed for the repeated simulation–optimization (SO) process in the management of coastal aquifers (Rao et al. 2003).

The surrogate technique is generally used to approximate the finite difference or finite element numerical code for solving density-dependent flow equations in a SO process (Forrester and Keane, 2009). Substantial data driven models such as ANN and modular neural networks (MNN) are reported as fast and popular surrogate models in SI management problems (Razavi et al. 2012; Bhattacharjya and Datta, 2005; Kourakos and Mantoglou, 2009).

There are a variety of other data driven techniques for being used as surrogate models in groundwater management problems subjected to SI such as Genetic Programming (Sreekanth and Datta, 2010), Evolutionary Polynomial Regression (Hussain et al. 2015), Radial Basis Functions (Christelis and Mantoglou, 2016a,b), Fuzzy c-mean clustering (Roy and Datta, 2017a,b) and M5 tree (Ranjbar and Mahjouri, 2018). Also, many machine learning techniques such as Fuzzy C-Mean Clustering (Ay and Kisi, 2014), Kriging and support vector regression techniques (Ouyang et al. 2017), iterative ensemble smoother (Chang et al. 2017), support vector machine (Alagha et al., 2017) and multivariate adaptive regression spline ensembles (En-MARS) have been extensively applied as a surrogate model in other SO problems (Roy

* Corresponding author.
E-mail addresses: aranjbar@mail.kntu.ac.ir (A. Ranjbar); mahjouri@kntu.ac.ir (N. Mahjouri); chrcld@unife.it (C. Cherubini)

https://doi.org/10.1016/j.jher.2019.11.005
Received 5 September 2018; Received in revised form 29 October 2019; Accepted 20 November 2019
Available online xxx

UNCORRECTED PROOF

and Datta, 2017a,b). However, the appropriateness of the mentioned techniques in terms of speed and accuracy should be examined for groundwater management problems subjected to SI using SO models, where SI is influenced by various uncertain decision variables. A review on the performance of surrogate models was carried out by Ketabchi and Ataie-Ashtiani (2015) which introduced Genetic Programming and ANN as accurate and fast tools used in such problems.

A comparison between MNN and Genetic Programming in terms of efficiency and robustness of SO solutions was made by Sreekanth and Datta (2010). It was found that by combining multi-objective Genetic Algorithm (MOGA) with Genetic Programming and MNN, only 2% of the runtime of the embedded FEMWATER-MOGA model is needed. It was also suggested that the GP can decrease the uncertainty of model predictions in optimization problems.

To the best knowledge of the authors, the efficiency of the conjunctive surrogate models with dynamic training has not been evaluated in saltwater intrusion problems. In addition, The present paper evaluates the performance of three machine learning-based algorithms (i.e. M5 tree, ANN and random subspace model (RSM)) as surrogate models for SEAWAT. The three models are trained using input–output samples provided by SEAWAT. The aquifer area is divided into five zones using k-means clustering technique. In order to enhance the accuracy of surrogate models, a conjunctive model using a linear function of the three models is developed. The conjunctive model is coupled with a multi-objective optimization algorithm, with the objectives of minimizing the SI length and maximizing the profit obtained from the agricultural zones, to find a Pareto-optimal front of solutions. To assess the efficiency of the conjunctive model, a comparison between the embedded SO model and the proposed surrogate models in terms of computational time and accuracy is carried out. Finally, the Nash bargaining theory is applied to select the stakeholders' preferred scenarios of utilizing the groundwater of Kahak aquifer adjacent to the Salt lake in Iran. The novel contributions of this study are as follows: 1) development of an efficient conjunctive surrogate model using a combination of three individual simulation models as surrogates for SEAWAT; 2) Developing a new simulation–optimization-based methodology for managing saltwater intrusion in aquifers with high number of pumping wells; (3) Selecting socially acceptable scenarios out of non-dominated solutions proposed by the optimization model coupled with the conjunctive surrogate model using the Nash bargaining theory; (4) Applying the methodology to a real aquifer suffering from saltwater intrusion from a saline lake (namely, the Salt lake).

2. Material and methods

In this paper, the structures of the three data-driven models of ANN, M5 tree and RSM, which are used as surrogate models for SEAWAT, are optimized. Furthermore, a conjunctive algorithm using a linear function of the three surrogate models is developed to improve the individual performance of surrogate models. The ANN model used in this paper is a two-layer perceptron. The architecture of the ANN is determined by trial and error. Also, the optimal number of nodes and depth of the M5 tree is obtained to predict the unseen data. In the RSM, different sub-spaces are generated to decrease the effect of repeated features. Finally, the statistical performance indices of the abovementioned models in calibration and validation phases are calculated and compared. After the development of an efficient conjunctive surrogate model, the structure of the multi-objective optimization model is presented.

2.1. M5 Tree model

Model trees are used to solve nonlinear problems by dividing the input area into sub-spaces and assigning a relationship to each sub-space (Bhattacharya et al. 2007). For each sub-space, there is one equation, which provides the outputs. The main preference of model trees to other simulation algorithms such as ANN is that they present understandable and linear relations. The M5 tree is one of the most popular tree-based models, which assigns a linear equation to each sub-space. The M5 tree splits the input area into many small spaces and fits the best regression model to each sub-space. The splitting process can be explained by a recursive moving from top of the tree to its leaves regarding decision roles as illustrated in Fig. 1a and Fig. 1b.

M5 tree algorithm computes the standard deviation reduction index (SDR) for partitioning protion T of data that reaches a node (Witten et al. 2006):

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{1}$$

where, T_i denotes i^{th} set after splitting T and sd is standard deviation function. The partitioning process is carried out for all nodes to create many branches. Even though the prediction error is usually reduced by increasing the number of branches, after reaching a specific number of branches, the error starts to increase again. This is called pruning, which happens when tree growth leads to over-training and the complex tree cannot necessarily predict the unseen data. Therefore, a stopping criterion should be considered to avoid growing a large tree. The pruning step uses the Gini index (Hastie et al. 2009) which determines the tree complexity corresponding to the generated error.

To have good predictions, the performance of the tree should not be dependent on the training data sets. If with a small change of the values in the training sets, the structure of the tree significantly changes, the tree cannot be considered as a smooth tree (Hastie et al. 2009). The smoothing criterion is to provide a tree that shows low sensitivity to training data. The M5 tree model used in this paper is a pruning M5 tree with optimized number of nodes for increasing the tree generalization.

2.2. Ann

In this paper, an ANN-based meta model is used as a surrogate model for SEWAT. In this study, the recharge rates of aquifer (R), pumping rates of five agricultural zones ($Q_1$to Q_5) and piezometric head near the salt lake boundary are inputs of the model while, saltwater intrusion length ($\bar{SI}$) and piezometric head after one month are outputs.

The ANNs are popular machine learning algorithms with structures like the behavior of neurons in the human brain. Performance of each neuron can be as follows (Ayoubloo et al. 2010):

$$y = f\sum_{j=0}^{n}(w_j x_j \tag{2}$$

where, y and x_j are output and input neurons, respectively, f is a nonlinear function and w_j represents the weights. Generally, for classification and training cases, f is considered as sigmoid function:

$$f(\theta) = \frac{1}{1+e^{-\theta}} \tag{3}$$

In this paper, an ANN based on multi-layer perceptron (MLP) algorithm is developed. To obtain an optimal structure for the ANN (the number of layers and nodes), a trial and error process is used (Fig. 1c). For each structure, the optimal values of weights are determined using the gradient of weights (Δw_i) by calculating the error between predicted (p_i) and target (t_i) values of instances as below:

$$E = \frac{1}{2}\sum (t_i - p_i)^2 \tag{4}$$

$$\Delta w_i = -\frac{dE}{dw_i} \tag{5}$$

where, dE and dw_i are the derivatives of error and weight, respectively. The value of Δw_i is multiplied by the learning rate and momentum coefficient for restricting the risk of obtaining a locally optimum solution.

2.3. Random subspace model (RSM)

In data-driven models, the random subspace algorithm is classified in the category of attribute- selection models where the error between predicted and

UNCORRECTED PROOF

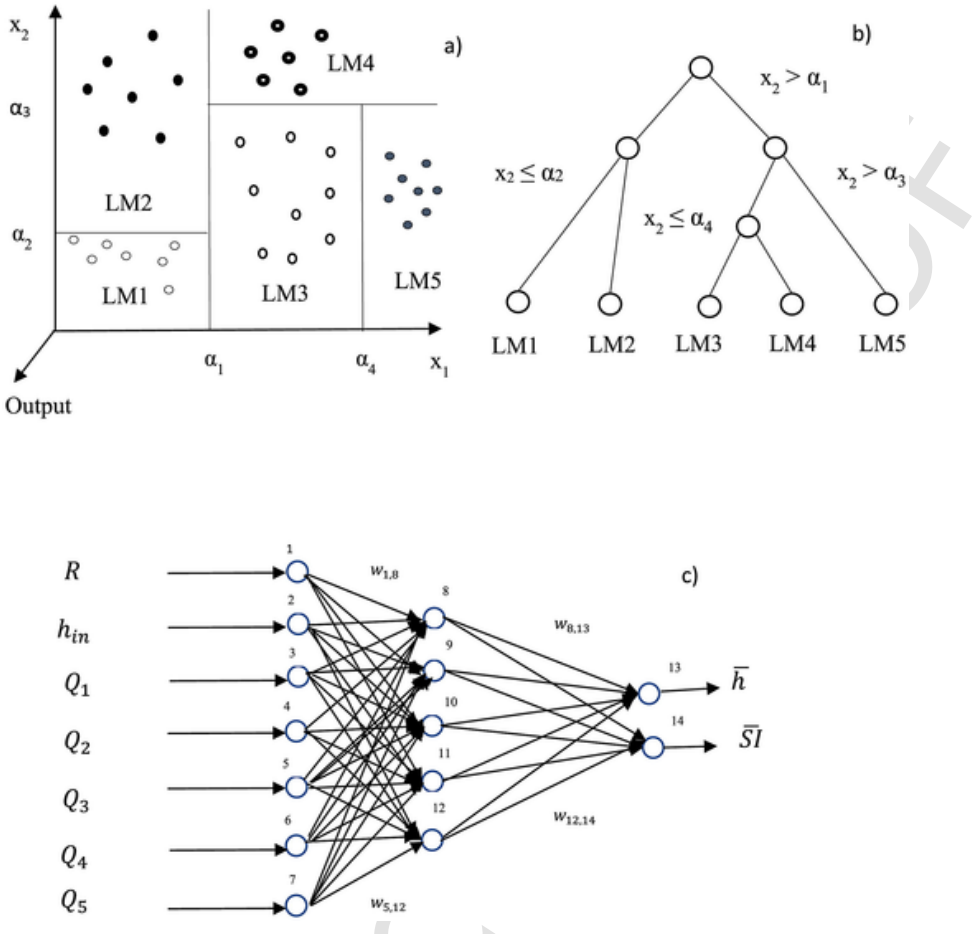


Fig. 1. Partitioning of input area and assigning linear models using M5 tree and ANN a) dividing input area and assigning regression line; b) prediction using generated linear relationships c) the structure of the optimal ANN.

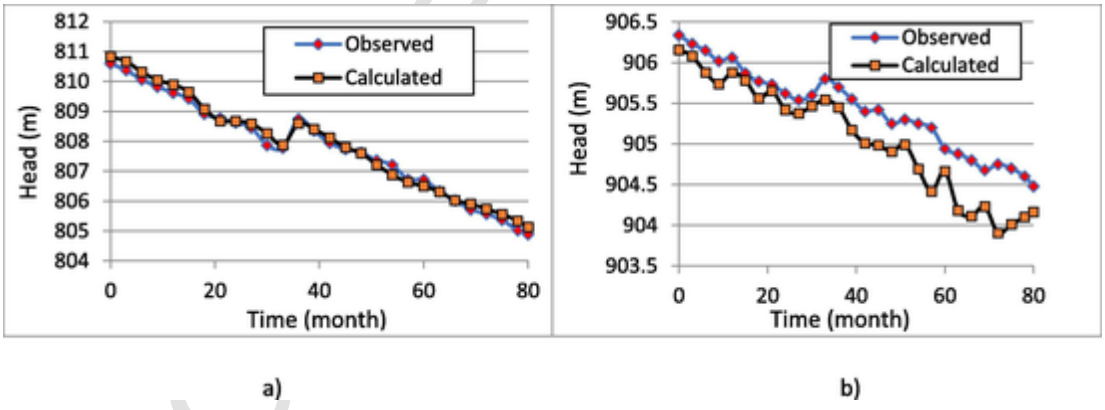


Fig. 2. Calculated and observed groundwater heads from 2006 to 2012 in the two nearest observation wells to the Salt lake: a) observation well 16, b) observation well 13.

observed data is decreased by dividing the features into many spaces (Breiman, 1996). The RSM algorithm is based on the fact that the average error of randomly partitioned features (input and output samples) is lower than that of a global training. We consider a prediction model using simple linear regression including M sub spaces and output (y) and input (x) data. A linear regression model in RSM approach versus simple regression can be written as follows:

$$y_{RSM}(x) = \frac{1}{M}\sum_{m=1}^{M} y_m(x) \tag{6}$$

$$y_m(x) = h(x) + e_m(x) \tag{7}$$

in which, $m = 1, \ldots, M$, which denotes the number of sub-spaces and $h(x)$ represents the predicted value using simple regression.

Additionally, the error of RSM (E_{RSM}) compared to average error of each sub space (E_{AV}) is decreased with increasing the number of sub spaces (M) as expressed below (Breiman, 1996):

$$E_{RSM} = \frac{1}{M}E_{AV} \tag{8}$$

Therefore, this algorithm is a good alternative for many complex models in which the number of attributes is more than the training data. In this paper, the features are ranked according to the correlation coefficient of their mean value for training samples.

2.4. Conjunctive surrogate models

In previous studies, conjunctive use of simulation models have been reported as useful tool for modeling non-linear simulation problems. A conjunctive surrogate mdoel combines the outputs of individual surrogate models by weighting them to enhance the simulations done by individual surrogate models (Ouyang et al., 2017, Hosseini and Kerachian, 2017). Hence, in this paper a conjunctive surrogate model is developed as a substitute for SEAWAT code. This conjunctive model is based on minimizing error criteria (E):

$$E_{con} = \frac{1}{k}\int_{k}^{0} (out\,(x) - out_{ens}\,(x))^2 dx = m^T R m \tag{9}$$

where, m denotes the weight of each surrogate model, $\mathrm{out}\,(x)$ and out_{con} are observed and predicted values of input x, respectively. R stands for the covariance matrix which are calculated as (Ouyang et al., 2017):

$$R_{ij} = \frac{1}{n}\alpha^T{}_i \alpha_j \tag{10}$$

where, n is the number of training instances, i and j represent different surrogate models and α is the vector of errors. The conjunctive model with the smallest discrepancy between predicted and observed values of training samples is selected as the best meta-model. Based on literature, the integration of ANN and M5 tree with other surrogate models will increase prediction accuracy (Ouyang et al., 2017; Yasa and Etemad-Shahidi, 2014). Moreover, the performance of weighted RSM for high dimensional data can be significantly improved even with several weaker models (Miraki et al., 2019). Therefore, in this paper, it is aimed to identify the optimal relative weights for the models of M5, ANN, and RSM to have a better combined surrogate model (conjunctive model). During the training and validating phases of the conjunctive model, the parameters of all individual models and their relative weights are optimized.

2.5. K-means Clustering (KMC)

K-means clustering technique is one of the most popular methods of clustering which has been implemented for classification of data sets regarding their feature vectors (MacQueen, 1967). In KMS, data points are partitioned into clusters based on their Euclidean distance from the center of each cluster. The number and radius of each cluster is determined randomly and data are assigned to the closest clusters. This iterative process continues to achieve the minimum cumulative error for all data sets. Therefore, the efficiency of KMC is a function of number of clusters and their radiuses (Milligan and Cooper, 1985). Suppose k is the number of clusters for a dataset $\{x_1, x_2, \ldots, x_n\}$in which each data is a vector with d dimensions $\{F_1, F_2, \ldots, F_d\}$. KMC tries to find the optimal value for k to minimize the cumulative error for each cluster as expressed bellow:

$$J = \sum_{i=1}^{k}\sum_{x_k \in G_i} \|x_k - c_i\|^2 \tag{11}$$

where, J represents Euclidean distance between ith sample and c_i (as the center of cluster $\mathrm{G_i}$). The values of $\mathrm{G_i}$ are determined using the membership values (u_{ij}) as below:

$$u_{ij} = \begin{Bmatrix} 1 if \|x_j - c_i\|^2 \le \|x_k - c_i\|^2; for k \ne i \\ 0 if \|x_j - c_i\|^2 \le \|x_k - c_i\|^2; for k \ne i \end{Bmatrix} \tag{12}$$

$$\left|G_i\right| = \sum_{j=1}^{n} u_{ij} \tag{13}$$

Then, u_{ij} with a binary value of 1 or 0 is used to determine the cluster center as below:

$$c_i = \frac{1}{\left|G_i\right|}\sum_{x_k \in G_j} x_k \tag{14}$$

$u_{ij} = 1$ or 0 indicate whether the sample is placed inside or outside the cluster, respectively. In this paper, K-means clustering is used to reduce the number of decision variables in the optimization model. The method used for clustering wells is based on the fact that salinity concentration in the observation wells depends on the distance from pumping wells (Kourakos and Mantoglou, 2009).

2.6. Multi-objective optimization algorithm

In this paper, an optimization model is developed using the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) (Deb et al. 2002) which is coupled with SEAWAT and surrogate models to find a Pareto front of non-dominated solutions for the optimization problem. The NSGA-II generates random values for decision variables and calculates corresponding values of the objective functions based on the groundwater simulations done by SEAWAT model. The value of decision variables are updated in the selection and crossover phases of the genetic algorithms. This repeated process of simulation–optimization continues until obtaining non-dominated solutions. The multi-objective optimization model can be formulated as follows:

$$\min \bar{SI} = \sum_{i=1}^{n} (C_i l_i \times A_i)/bl_{mean} n C_{\max} \tag{15}$$

$$\max Q_t = \sum_{i=1}^{m} q_i \tag{16}$$

subject to:

$C_i < C_{\max}$

$q_i < q_{\max}$

in which, $\bar{SI}$ represents saltwater intrusion length or the location of saltwater wedge considering 50% *iso*-concentration profile, $n = 24$ and $m = 69$ are the number of observation and pumping wells, respectively, C_i is salinity concentration (Total Dissolved Solids (TDS)) in cell i, A_i is the area of ith cell, q_i is the pumping rate of ith well which is located in distance $l_i$from the coastline, b represents the mean length of the coastline, l_{mean} represent the mean distance of wells from the coastline and $2500\mathrm{mg/L} < C_{\max} \le 11000\mathrm{mg/L}$ and $q_{\max} = 2700 m^3/day$ are considered as the maximum allowed salt concentration and pumping rate, respectively. The maximum allowable TDS concentration varies with the crop type, which is divided into four classes $\left(C_{max1}, C_{max2}, C_{max3}, C_{max4}\right)$. The salinity class $C_{\max}$ denotes TDS concentration < 3000 mg/L, C_{max2} denotes TDS concentration between 2500 mg/L and 4000 mg/L, denotes C_{max3} denotes TDS concentration between 4000 mg/L and 7000 mg/L and C_{max4} denotes TDS concentration more than 7000 mg/L and 11000 mg/L for C_{max4}.

2.7. Conflict resolution

In order to incorporate the conflicting utilities of the stakeholders (herein, the agricultural sectors), the Nash bargaining theory is utilized. Using this theory, agricultural sectors aim at optimally sharing the total benefit obtained from

UNCORRECTED PROOF

exploiting the aquifer. The stakeholders from agricultural sectors intend to select the mostly approved scenario according to the gained profit (Nikoo et al. 2016). If the feasible region of solutions is bounded and convex, the Nash problem has a unique answer. Considering the above-mentioned criteria, the Nash solution can be obtained as follows (Nash 1950):

$$Max \quad M = (o_1 - e_1) \times (o_2 - e_2) \times \ldots \times (o_I - e_I), o_i \geq e_i, i = 1, \ldots, I \tag{17}$$

where, o_i is the profit of ith stakeholder, e_i is the utility of ith stakeholder before bargaining, I is the number of stakeholders and M represents the value of Nash function. For all scenarios on the Pareto front, M is calculated and the scenario with minimum value of M is selected as the Nash bargaining solution.

2.8. Study area

The capability of the developed methodology is examined on an unconfined aquifer, namely Kahak, located in Qom County near the Salt Lake in Iran. The main part of the geological sediment in the Kahak aquifer is permeable alluvial which is composed of conglomerates and fluvial terraces. Over-exploiting the groundwater for agricultural uses has led to saltwater intrusion to the aquifer. The location and the boundary conditions of the aquifer are shown in Fig. 3. As seen in this figure, the eastern part of the aquifer is mostly affected by saltwater with a concentration of the total dissolved solids (TDS) of about 16,000 mg/l and most of the east and west boundaries are impervious. The recharge rate is non-uniform and varies from 0.000063 to 0.000078 m/day. However, the sensitivity analysis on the SI length indicates that 69 pumping wells, out of 1565, located adjacent to the Salt Lake dominantly affect the SI (Ranjbar and Mahjouri, 2018, 2019).

Regarding a difference of 10 m between the upper and lower topographic elevations, the SI is influenced only by the pumping activity of the mentioned zone. Fig. 3 illustrates the computed contours of TDS concentration for the present abstraction rates obtained using variable density flow simulation model (i.e. SEAWAT). The 69 effective pumping wells are illustrated in Fig. 3.

The sensitivity analysis has shown that salinity distribution in the observation wells depends mostly on pumping from nearby wells rather than from distant ones (Kourakos and Mantoglou, 2009). A sensitivity analysis on the effect of pumping on TDS concentration in the observation wells is also implemented (see below Fig. 4). The TDS concentrations near the observation wells and mostly affected pumping wells adjacent to the observation wells are considered as inputs and outputs of the SEAWAT simulation model, respectively. The locations of pumping wells and observation wells are shown in Fig. 3. The wells that have a considerable effect on TDS concentration (i. e. more than 5%) in a specific observation point are considered as sensitive wells.

3. Development of the models

3.1. Calibration of the numerical simulation model

The study area with dimensions of 15 km (length) × 8 km (width) is discretized into 500 m × 500 m cells. Considering the geological and soil characteristics, the study area is discretized into 5 layers with different thicknesses. A simulation period of four years is considered which is divided into 48 monthly time steps. The initial TDS concentration in the northern boundaries near the Salt Lake is 12.5 (gr/L). The northern boundary (near the Salt Lake) is simulated using a time variant specified head considering hydraulic heads between 8 m and 12 m which vary with time. A 3rd order total variation diminishing (TVD) numerical scheme is used to solve the advection and dispersion equations for the TDS concentration in the SEAWAT model. The coefficients of hydraulic conductivity and specific yield of the groundwater are calibrated based on the transient condition in the simulation model. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between simulated and observed groundwater heads in 24 observation wells are minimized. Results show an acceptable accuracy in the validation phase after 2400 days, where the mean values of MAE and RMSE in the two nearest observations wells to the Salt lake at the end of the planning horizon (after 2400 days) are 1.52 m and 1.85 m, respectively (see Fig. 2a and b). Fig. 3 illustrates the salinity contours in the aquifer for September 2012.

3.2. Training process

The surrogate models are utilized to approximate the response of the SEAWAT model to different pumping and recharge conditions. The average value of groundwater head ($\overline{h}$) and average salinity concentration ($\overline{SI}$) in four observation wells near the Salt lake are considered as the outputs of the surrogate models while the rates of discharge from 69 pumping wells and the rates of recharge are the main inputs. The outputs of the surrogate models are defined

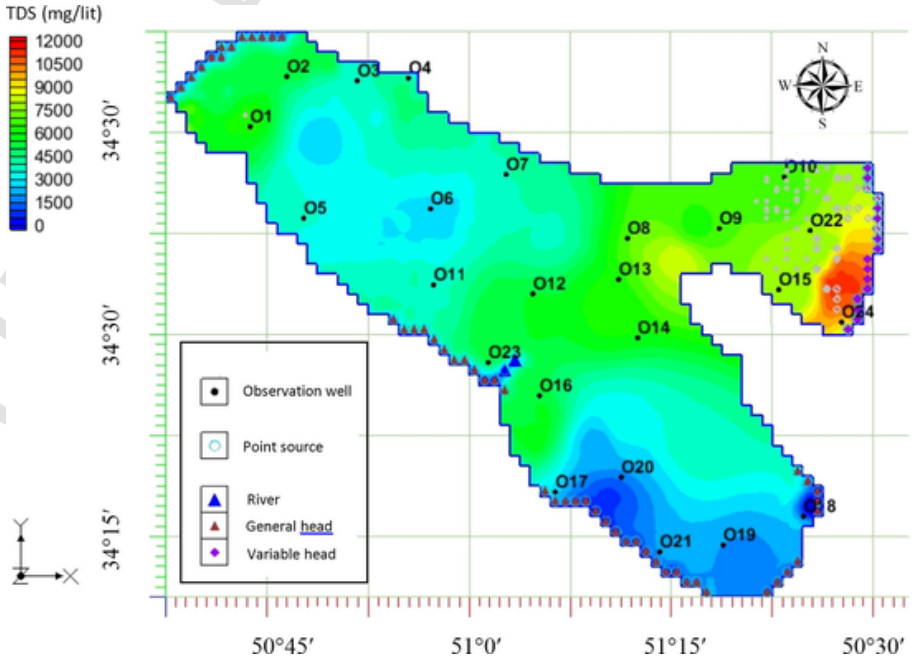


Fig. 3. Locations of the observation and pumping wells and the distribution of TDS concentration based on the current pumping rates.

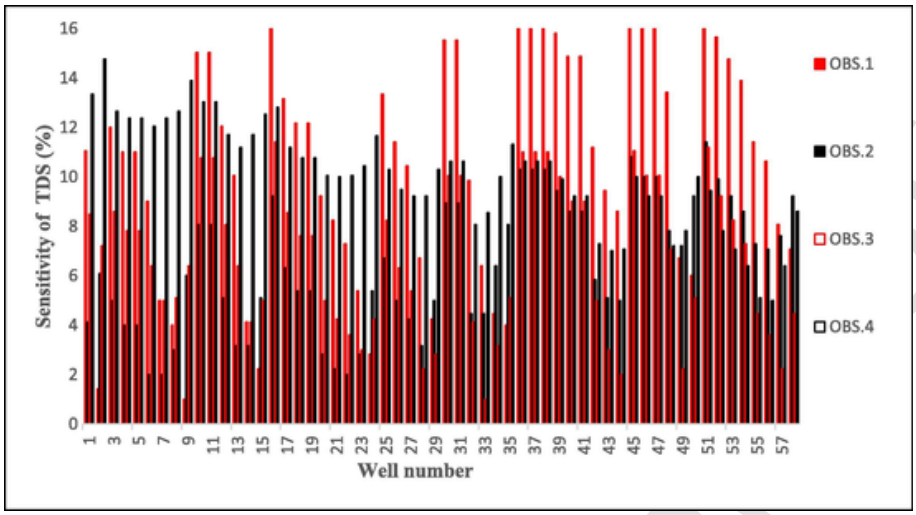


Fig. 4. Sensitivity of TDS concentration in the observation wells to the pumping rate in the pumping wells.

as follows:

$$\bar{SI} = \frac{\sum_{i=1}^{n} C_{(TDS)i}}{n} \tag{18}$$

$$\bar{h} = \frac{\sum_{i=1}^{n} h_i}{n} \tag{19}$$

where, $n = 4$ is the number of observation wells parallel to the shoreline, and C_{TDS} and h are salinity concentration and groundwater head in the ith well, respectively. The inputs include the abstraction rate from 69 active pumping wells (Q), average recharge rate (R), inland head (h_{in}) and salinity concentration at the beginning of a time step (three months) ($\bar{SI}_{in}$). The input–output data sets consist of 998 instances with a random and relatively Gaussian distribution. In the surrogate models, 75% (7 4 8) of the samples are considered for the training phase and 25% of samples are considered for the validation phase. To avoid overfitting to the observations, the training phase will finish when the error in the testing phase starts to increase. The ranges of input and output samples are presented in Table 1. The values of the calibrated parameters of the SEAWAT model are shown in Table 2.

3.3. Development of the surrogate models

Three surrogate models are developed to approximate simulations done using the SEAWAT model. Due to the large number of pumping wells, the aquifer area is divided into several clusters and the wells is each cluster are grouped. To do this, k-means clustering (KMC) technique is used which classifies the wells based on their Euclidean distances from the nearest observation well. Each surrogate model calculates the head and the TDS concentration in the observation wells located inside each cluster. Then, total pumping from a cluster is concentrated in an equivalent distance (ES) (m) calculated as below:

Table 1
Range of variables considered for the surrogate models.

Parameter	Range
$Q_{total}(m^3/day)$	200 – 68,400
$R(m/day)$	0.00001–0.000092
$h_{in}(m)$	690–806
$\bar{SI}(m)$	105–620
$\bar{h}(m)$	680–720

Table 2
The calibrated values of aquifer parameters using the SEAWAT model (Ranjbar and Mahjouri, 2018).

Parameter	Symbol	Value	Unit
Longitudinal hydraulic conductivities	K_{xx}	1 – 50	m/day
Horizontal anisotropy	λ_{xy}	1	m/day
Vertical anisotropy	λ_{xz}	8	m/day
Specific yield	S_y	0.05	–
Porosity	φ	0.31	–
Longitudinal dispersivity	α_L	7	m
Transverse dispersivity	α_T	1.5	m
Density difference ratio	ε	0.025	–
Vertical recharge rate	V_r	0.02	m/day
Reference hydraulic head	h_f	1.65	m
Molecular diffusion	d_0	6×10^{-7}	m^2/s

$$ES = \frac{\sum_{i=1}^{n} Q_i \times Ln(r_i)}{\sum_{i=1}^{n} Q_i} \tag{20}$$

where, $Q_i$and $r_i$are pumping rate and distance of ith well from the cluster center and n is the number of wells located in the cluster. The pumping and observation wells near the Salt Lake are illustrated in Fig. 3 The optimum value of k is calculated based on the location and pumping rates of 69 pumping wells as presented in Table 3. The performance of KMC is investigated using the silhouette graph (Rousseeuw, 1987).

The silhouette graph approach is used for interpretation and validation of clusters of samples. The number of clusters is calculated by silhouette graph approach in MATLAB software and shown in Fig. 5a. This graph is based on the minimum Euclidean distance obtained in iteration 20 as indicated in Fig. 5b. As illustrated in Fig. 3a, more samples in five clusters have a great silhouette value (i.e. 0.6).

After clustering the solution area of decision variables, a surrogate model is applied to the cluster. Regarding the large value of Q compared to other input variables, the logarithm of Q is used for training the M5 tree. The splitting of nodes is done only for samples with the size of 6% of total samples and more. This value for the bottom branch is considered as 2% of total samples. The number of nodes and depth of the tree is determined using trial and error to achieve the best performance corresponding to the test data. The structure of the M5 tree is shown in Fig. 6, in which the parameters and relations are illustrated in nodes and rectangles, respectively. The relations between $\bar{SI}$ and input parameters are defined using linear models (LMs). As shown in Fig. 6, the

UNCORRECTED PROOF

Table 3
Basic range of the data for each cluster.

Well number	Q (m^3/day)	r (m)■	Silhouette Value◇	Corresponding Cluster
1	472	1500	0.555918	5
2	715	3000	0.368701	5
3	787	3500	0.385491	5
4	235	6000	0.506376	2
5	1252	3000	0.470217	5
6	214	6000	0.505774	2
7	2093	2000	0.22874	5
8	2258	3500	0.450941	5
9	959	4500	0.554362	4
10	948	5000	0.71977	4
11	508	5500	0.130177	4
12	510	7000	0.651385	2
13	1080	1500	0.555918	5
67	0.97	10,000	0.644665	1
68	515	10,000	0.702108	1
69	515	10,000	0.702108	1

■ The Euclidian distance between a well and the nearest cluster center

value of saltwater intrusion length at the end of each time step ($\overline{SI}$) is a function of piezometirc head in observation wells near the salt lake (H_m), saltwater intrusion at the beginning of time step (SI) and pumping rate from five agricultural zones (i.e. Serajeh (Q_s), Noran (Q_n), Malekan (Q_{ma}), Dolatabad (Q_d) and Momenabad (Q_m)).

Hornik et al (1989) recommended that for engineering problems, ANN with a hidden layer can show the best performance. In this paper, a two-layer perceptron is selected and the number of nodes in each layer is determined through trial and error. The Levenberg-Marquardt training algorithm with learning rate of 0.4 shows the lowest discrepancy between the predicted and simulated data. The training process is finished when the number of epochs reaches 1500, or the error becomes less than a threshold value. The optimal weights and the structure of the ANN are presented in Table 4.

Moreover, a random subspace model (RSM) is developed for prediction corresponding to the 8 input variables. The size of 11 trees generated using a random subspace varies between 59 and 163, and the total number of nodes is 7080. As an example, the proposed $\overline{SI}$ by tree 1 and subspace 42 for the specific range of input variables is shown in Table 5.

In order to improve the efficiency of the surrogate models, a conjunctive algorithm using a combination of the predictions by the three models is evaluated. The conjunctive model is defined as a linear function of M5, ANN, and RSM predictions and can be written as follows:

$$P_C = \frac{K_1 P_{M5} + K_2 P_{ANN} + K_3 P_{RSM}}{K_1 + K_2 + K_3} \quad (21)$$

where, P_C denotes prediction by the conjunctive model, P_{M5}, P_{ANN} and P_{RSM} stand for predictions by the individual surrogate models of M5, ANN and RSM, respectively and K_1, K_2 and K_3 are constant coefficients which show the relative weights of estimations by individual models. To develop the conjunctive *meta*-model, the general structure of the M5, ANN, and RSM simulation models are assumed to be similar to those obtained when training them individually. The optimal values of the parameters of all individual models as well as K_1, K_2 and K_3 coefficients are obtained through a one-leave out cross validation process.

3.4. Optimization process

The conjunctive surrogate model (CM) which has been trained and validated using the outputs of the SEAWAT model is linked with NSGAII optimization model to find the optimum groundwater withdrawal scenarios through a 20-year planning horizon. Through the optimization process, the SEAWAT model generates new values for the decision variables (pumping rates from 69 wells) to update the surrogate models. This repeated process is terminated when the values of decision variables vary slowly. Also, to increase the accuracy, the numerical model generates new samples around the global optimum value. In the optimization model, the population size and the number of generations are set to be 60 and 150, respectively. The population size in NSGAII is 400. Mutation and cross over coefficient are set as 90 and 55 percent, respectively. To improve the efficiency of the selection step and maintain the diversity of solutions, the crowding distance method (Deb et al. 2002) is applied. Each solution is mutated by inserting Gaussian distribution to create near solutions. This technique tries to replace the current solution with neighboring solution. The values of input variables are between 150 and 1620 (m^3/day) which vary with seasons. The training data (998 samples) have a Gaussian distribution. As seen in Fig. 7, the surrogate models are retrained using the input–output data of the numerical model to improve its accuracy. Also, the surrogate models are trained for different clustering schemes to find the best number of clusters. Moreover, the optimization algorithm is updated using the results of the surrogate models to increase the number of training samples around the optimum solution. If the convergence criterion is satisfied, a Pareto font of optimal solutions is derived and the value of the Nash product is calculated for each Pareto-optimal solution (scenario of groundwater withdrawal).

3.5. Performance criteria

To evaluate the performance of the three mentioned surrogate models, 198 testing datasets are selected. The performance of the models is evaluated using the correlation coefficient (CC) and the mean squared error (MSE) and the

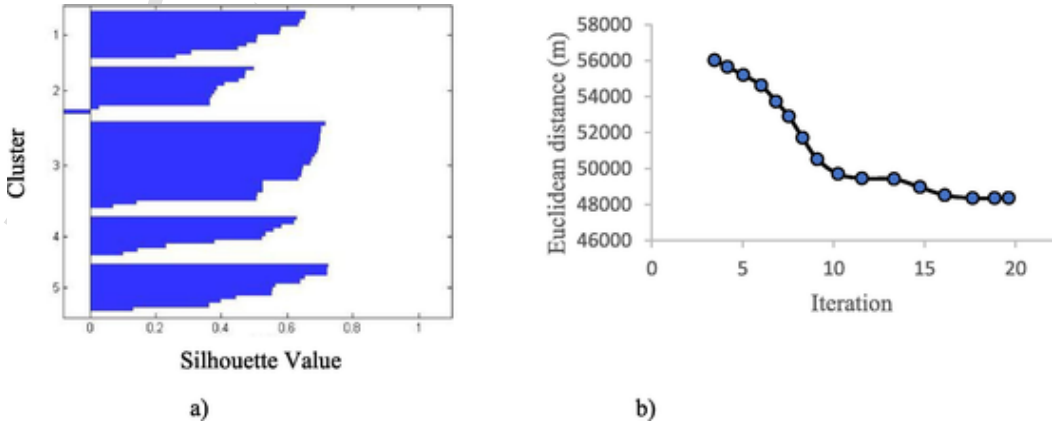
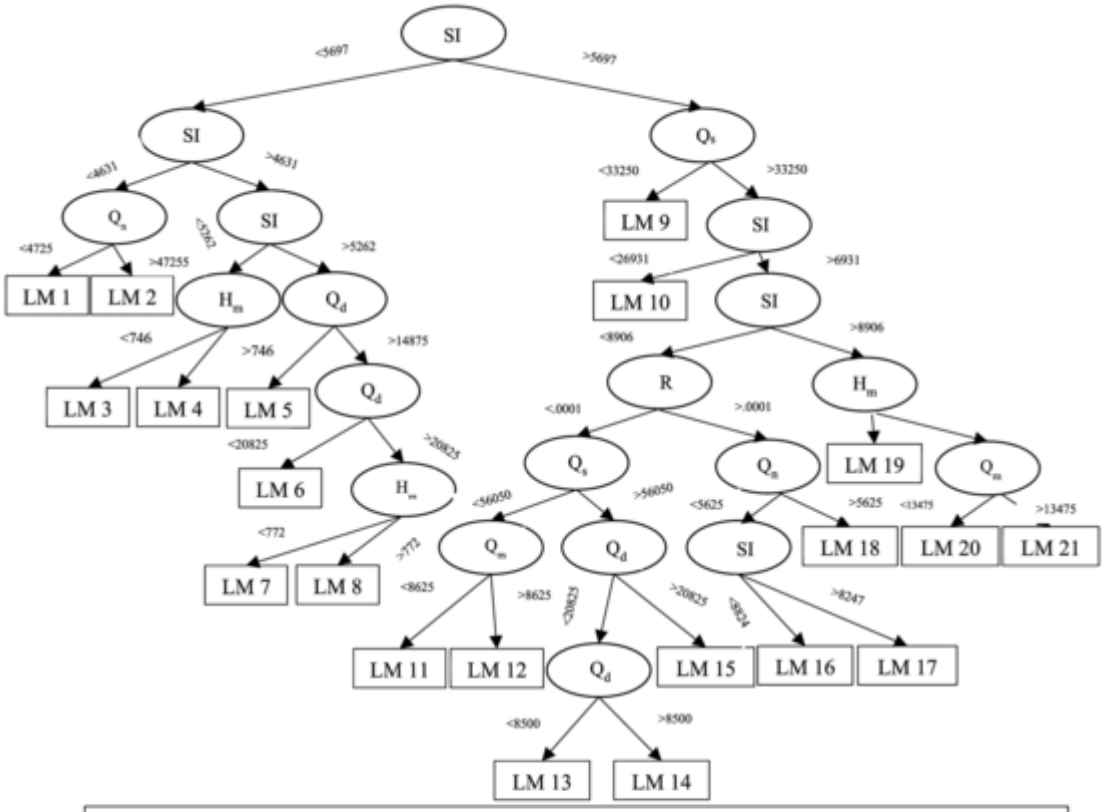


Fig. 5. Silhouette graph for the five clusters corresponding to the final iteration a) Silhouette value versus the number of clusters b) convergence to a minimum Euclidean distance.

CORRECTED PROOF



$$\text{LM 1:}\quad \frac{\overline{SI}}{76540} = e^{-0.073\ln(Q_m)-0.019\ln(Q_{ma})+0.071\ln(Q_s)-0.108\ln(Q_n)-0.006\,(\mathrm{R})-0.39\,(H_m)-0.196\,(\mathrm{SI})}$$

$$\text{LM 2:}\quad \frac{\overline{SI}}{632.8} = e^{-0.088\ln(Q_m)-0.089\ln(Q_d)+0.128\ln(Q_{ma})+0.012\ln(Q_s)-0.003\ln(Q_n)-0.006\,(\mathrm{R})-0.134\,(H_m)+0.046\,(\mathrm{SI})}$$

$$\text{LM 3:}\quad \frac{\overline{SI}}{384.4} = e^{-0.250\ln(Q_m)+0.247\ln(Q_{ma})-0.011\ln(Q_s)-0.003\ln(Q_n)-0.0002\,(\mathrm{R})-0.133\,(H_m)+0.046\,(\mathrm{SI})}$$

$$\text{LM 4:}\quad \frac{\overline{SI}}{384.4} = e^{-0.220\ln(Q_m)+0.247\ln(Q_{ma})-0.009\ln(Q_s)-0.003\ln(Q_n)-0.0162\,(\mathrm{R})-0.103\,(H_m)+0.046\,(\mathrm{SI})}$$

.

.

$$\text{LM 20:}\quad \frac{\overline{SI}}{341.6} = e^{+0.337\ln(Q_m)-0.255\ln(Q_{ma})-0.018\ln(Q_s)-0.007\ln(Q_n)-0.0001\,(\mathrm{R})-0.133\,(H_m)+0.270\,(\mathrm{SI})}$$

$$\text{LM 21:}\quad \frac{\overline{SI}}{67.8} = e^{+0.401\ln(Q_m)-0.255\ln(Q_{ma})-0.018\ln(Q_s)-0.007\ln(Q_n)-0.001\,(\mathrm{R})-0.133\,(H_m)+0.165\,(\mathrm{SI})}$$

Fig. 6. The relations presented by MM5 tree for $\overline{SI}$ considering different ranges for dimensionless inputs.

Table 4
The values of parameters of the optimum ANN.

Weights				Biases
$w_{1,8}$ = 0.32	$w_{2,8}$ = 0.61	…	$w_{8,14}$ = -2.12	1.45
$w_{1,9}$ = -0.37	$w_{2,9}$ = 5.42	…	$w_{9,14}$ = -1.29	4.35
$w_{1,10}$ = -0.14	$w_{2,10}$ = -0.02	…	$w_{10,14}$ = -1.62	−2.12
$w_{1,11}$ = 1.17	$w_{2,11}$ = 1.62	…	$w_{11,14}$ = 0.049	− 3.07
$w_{1,12}$ = -0.42	$w_{2,12}$ = 0.049	…	$w_{12,14}$ = 0.14	−3.74

mean of absolute errors (ME) between predicted and observed values as follows:

$$CC = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(s_i-\bar{s})}{\left(\sum_{i=1}^{n}(p_i-\bar{p})^2\sum_{i=1}^{n}(s_i-\bar{s})^2\right)^{0.5}} \tag{22}$$

$$MSE = \frac{\sum_{i=1}^{n}(s_i-p_i)^2}{n} \tag{23}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_i}{s_i} \tag{24}$$

in which, p_i and s_i are respectively predicted and simulated values for ith sam-

Table 5
The values of $\overline{SI}$ corresponding to random tree 1 and space 42.

$\overline{SI}$ =317.47	$\overline{SI}$ =223.77	$\overline{SI}$ =171.1	$\overline{SI}$ =238.17	$\overline{SI}$ = 226.25	$\overline{SI}$ =
$Q_5 < 7875$	$Q_1 \geq 8525$	$Q_4 \geq 46550$	$Q_1 \geq 14025$	$Q_5 \geq 7875$	$Q_1 \geq$
$Q_1 < 14025$	$Q_4 < 2025$	$Q_4 < 61750$	$Q_4 < 33250$	$Q_1 < 15125$	$Q_5 <$
$Q_2 < 4675$	$Q_4 < 46550$	$Q_5 < 675$	$Q_3 < 21235$	$Q_1 < 3575$	$Q_1 <$
$R < 0.000012$ $h_{in} < 780$	$R < 0.000016$ $h_{in} < 763$	$Q_3 < 8552$ h_{in} greater than 785	R greater than 0.000056	$Q_5 < 12825$ $Q_2 < 21675$	h_{in} gre than R gre than (

ple, $\overline{p}$ and $\overline{s}$ represent mean values for the predicted and simulated data, respectively and n is the number of testing samples.

4. Results and discussion

4.1. Prediction of the SI length

Based on Figs. 8 and 9, an acceptable match can be seen between the results of the ANN, M5 and RSM and those of the SEAWAT. Moreover, the values of the statistical criteria in the validation phase using 198 samples for the three surrogate models are presented in Table 6.

Among the three surrogate models, M5 has the highest value of CC and the lowest MSE in the training phase. Based on Fig. 9, for larger values of $\overline{SI}$, two models of RSM and M5 respectively underestimate and overestimate $\overline{SI}$. Also, ME and MSE indices increase with increasing $\overline{SI}$. However, for $\overline{SI} \leq 320$ m and $\overline{h} \leq 750$ m, the results of RSM are more accurate than those of the other models. M5 and ANN models show more accuracy in predicting $\overline{SI}$ and $\overline{h}$values greater than 320 m and 750 m, respectively. Generally, it can be concluded that the ANN results are more accurate for data which are not in the training samples and hence, the ANN model has a better generalization ability. To develop the conjunctive model (CM), the structure of the three surrogate models individual are considered to be fixed but the values of their parameters and relative weights are optimized through a one-leave-out cross validation process. For the conjunctive model, the optimal values of K_1, K_2 and K_3 are determined as 0.4, 0.3 and 0.3 (m/day), respectively.

As shown in Figs. 10 and 11 and Table 6, the CC values for SI and $\overline{h}$ prediction in the validation phase of the CM are 0.98 and 0.97, respectively. Additionally, the conjunctive model has lower values of error indices (i.e. MSE and ME of about 32%) comparing to the best single surrogate model (i. e. M5 model) and can be selected as the final surrogate model.

4.2. Management of groundwater extraction

Fig. 12a illustrates the obtained Pareto-optimal front considering the two objective functions of for SEAWT-NSGAII and CM-NSGAII. The horizontal axis of Pareto front shows the total discharge rates from the pumping wells and the vertical axis shows the saltwater intrusion length ($\overline{SI}$) at the end of the 20-year planning horizon. As seen, the slope of trade-off curve decreases with increase of pumping rates. According to Fig. 12, the maximum groundwater extraction is suggested to be from wells far from the shoreline and near the optimum locations. The results show that the trade-off curves generated by the two

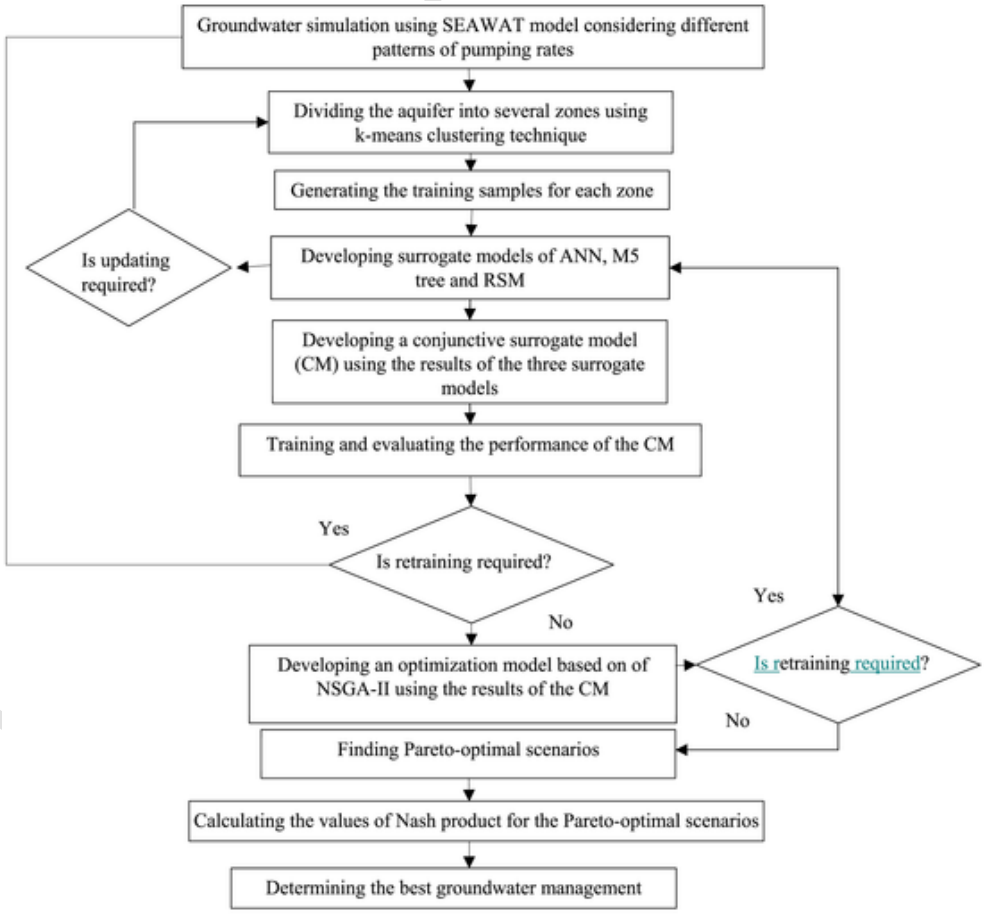


Fig. 7. The flowchart of the proposed methodology for management of aquifers subjected to saltwater intrusion.

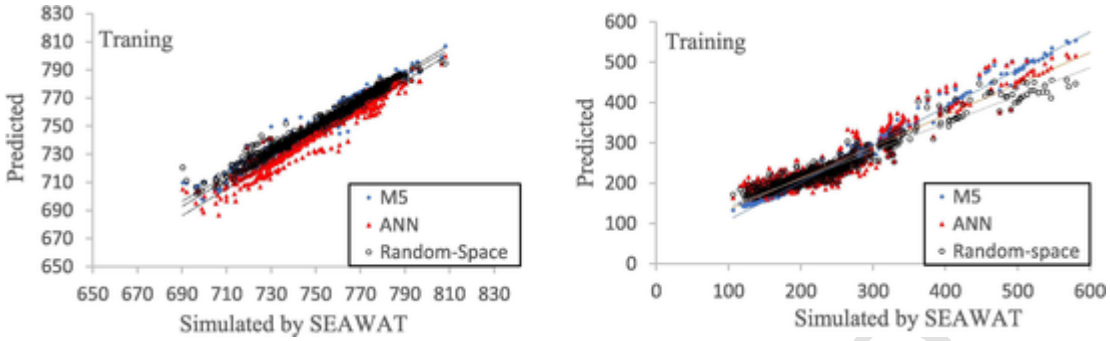


Fig. 8. The simulated and predicted $\bar{SI}$ (m) using the three surrogate models in the training phase.

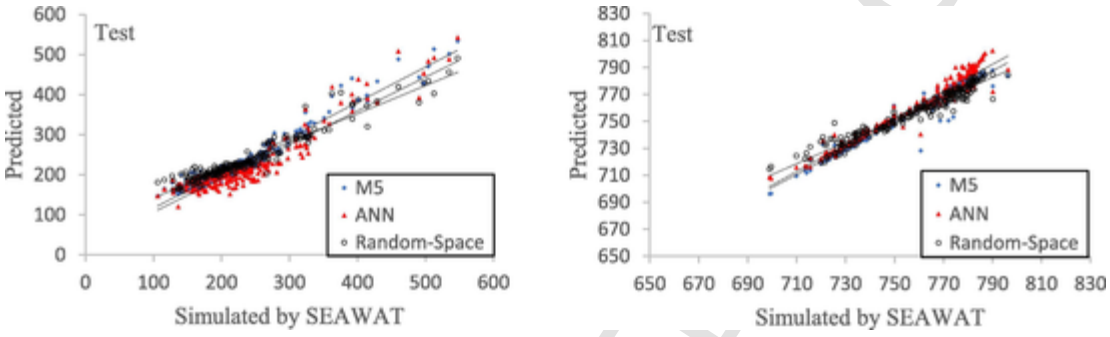


Fig. 9. The simulated and predicted $\bar{h}$ (m) using the three surrogate models in the testing phase.

Table 6
The values of the statistical indices for the three surrogate models and the conjunctive model (CM).

Models	CC ($\bar{SI}$)	CC ($\bar{h}$)	MSE ($\bar{SI}$)	MSE ($\bar{h}$)	ME ($\bar{SI}$)	ME ($\bar{h}$)
M5 (training)	0.96	0.96	484	30.25	16.54	2.98
ANN (training)	0.89	0.93	1369	222	29.7	13.5
RSM (training)	0.94	0.95	1602	46.24	27.65	4.45
CM (training)	0.99	0.99	377	22.50	12.46	2.29
M5 (test)	0.92	0.95	961	34.45	22.54	2.99
ANN (test)	0.88	0.92	1346	40.96	29.07	4.13
RSM (test)	0.84	0.92	2704	88.33	37.2	6.98
CM (test)	0.98	0.97	730	25.84	17.6	2.43

SO models are relatively similar. However, for a limited number of scenarios, the results of CM-NSGAII are relatively overestimated. Additionally, for scenarios which suggest high pumping rates, the discrepancy between the results of the two approaches is low. The required time for converging to optimum solutions, using CM-NSGAII model on a PC with a configuration of Intel Core™ 7 (considering 400 populations), is presented in Table 7. As seen in this table, the computational time required for the CM-NSGAII model is only 5% of the time that SEAWAT-NAGA-II requires to converge.

Also, for CM-NSGA-II, the major portion of the runtime is related to the simulation using the SEAWAT code and the time used for generating input–output datasets. During generating the datasets for the three surrogate models, SEAWAT saves the data in a HDF5-format file for MATLAB and this process increases the runtime of the SO model. Overall, the multi-objective optimization algorithm using a combination of the conjunctive surrogate model (CM) and NAGA-II significantly decreases the computational burden in the problem of management of an aquifer under saltwater intrusion.

4.3. Conflict resolution

As seen in Fig. 3, water withdrawal from wells near the Salt Lake (i.e. Momen and Dolat) has a significant impact on saltwater intrusion. The values of the Nash product or Nash function (F value) for Pareto-optimal scenarios are shown in Fig. 12b. According to this figure, Scenario 4 with F value of 0.0128 is the most approved scenario by the agricultural sectors. The total groundwater pumping rate of five agricultural zones corresponding to Scenario 4 are 132, 77, 115, 838 and 458 million cubic meters over 20 years, respectively. Regarding the allowed rate of extraction (about 1600 m^3/day) for each zone, it can be

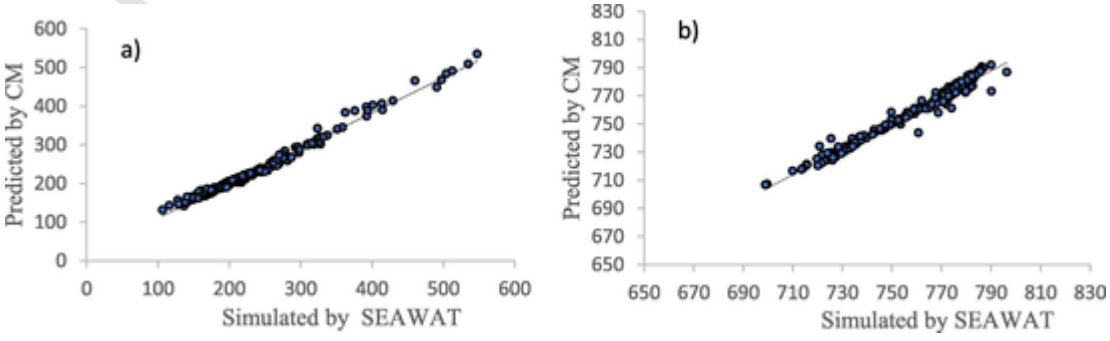


Fig. 10. The simulated and predicted values of a) $\bar{SI}$ and b) $\bar{h}$(m) using the conjunctive model (CM) for 120 validation data.

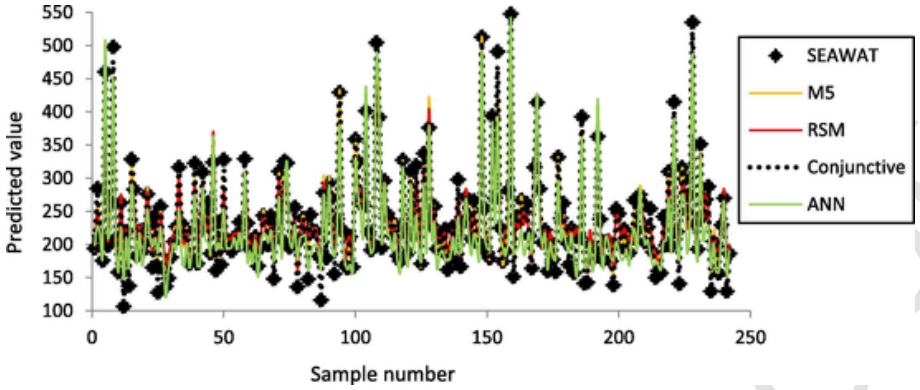


Fig. 11. Variation of the simulated $\overline{SI}$(m) using the models of M5, ANN, RSM, conjunctive and SEAWAT based on 250 validation data.

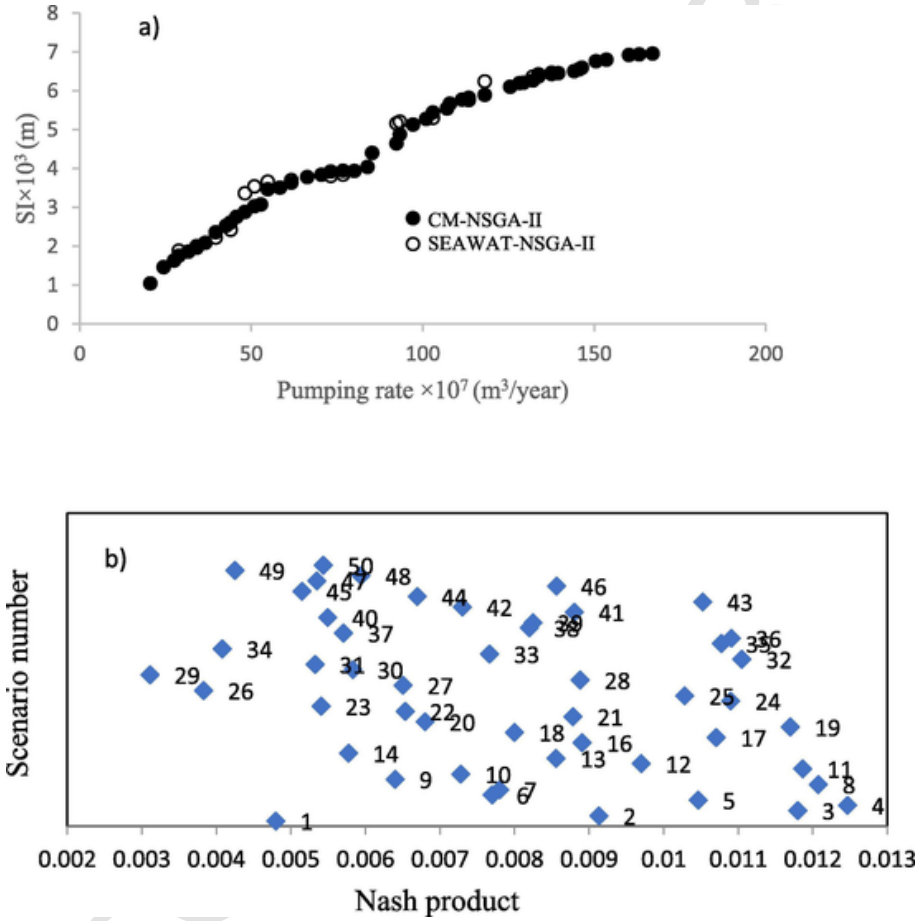


Fig. 12. a) The obtained trade-off curve for the management solutions using CM-NSGA-II and SEWAT-NSGA-II, b) The values of Nash product (F) corresponding to the optimal scenarios.

Table 7
Comparison of computational performance for different simulation–optimization (SO) model.

SO model	SEAWAT model calls	CM calls	Computational time (h)
SEAWAT-NAGAII	15,100	0	294
CM-NAGAII	280	14,200	14.2

concluded that at least 50% of the water demands of the agricultural zones is supplied.

Also, the value of $\overline{SI}$ for Scenario 4 is 6550 m that is about 87% of the maximum $\overline{SI}$ (7500 m). However, the best scenario is selected considering $\overline{SI}$ and profit simultaneously. For this purpose, four scenarios with large values of Nash product are evaluated in term of SI. Fig. 13a illustrates the TDS contours ranging from about 1000 mg/L to 15000 mg/L corresponding to four scenarios with a the highest Nash values. The distribution of the TDS for the agricultural zones is illustrated in Fig. 13a. As seen in Fig. 13b, the SI length for F = 0.011 is about 5300 m (see Table 8).

Interestingly, Fig. 13c with F = 0.0124 shows a bigger $\overline{SI}$. However, for the case with F = 0.0121, $\overline{SI}$ is 7800 m. According to the TDS contours for the four scenarios, it can be concluded that the most polluted area is located in Zone 4 where the TDS concentration is more than 8000 mg/L. Therefore, the abstraction rate in this area should be decreased to about 50%. As presented in Table 8, Scenario 2 with F = 0.0124 and SI = 6400 m has the lowest pumping

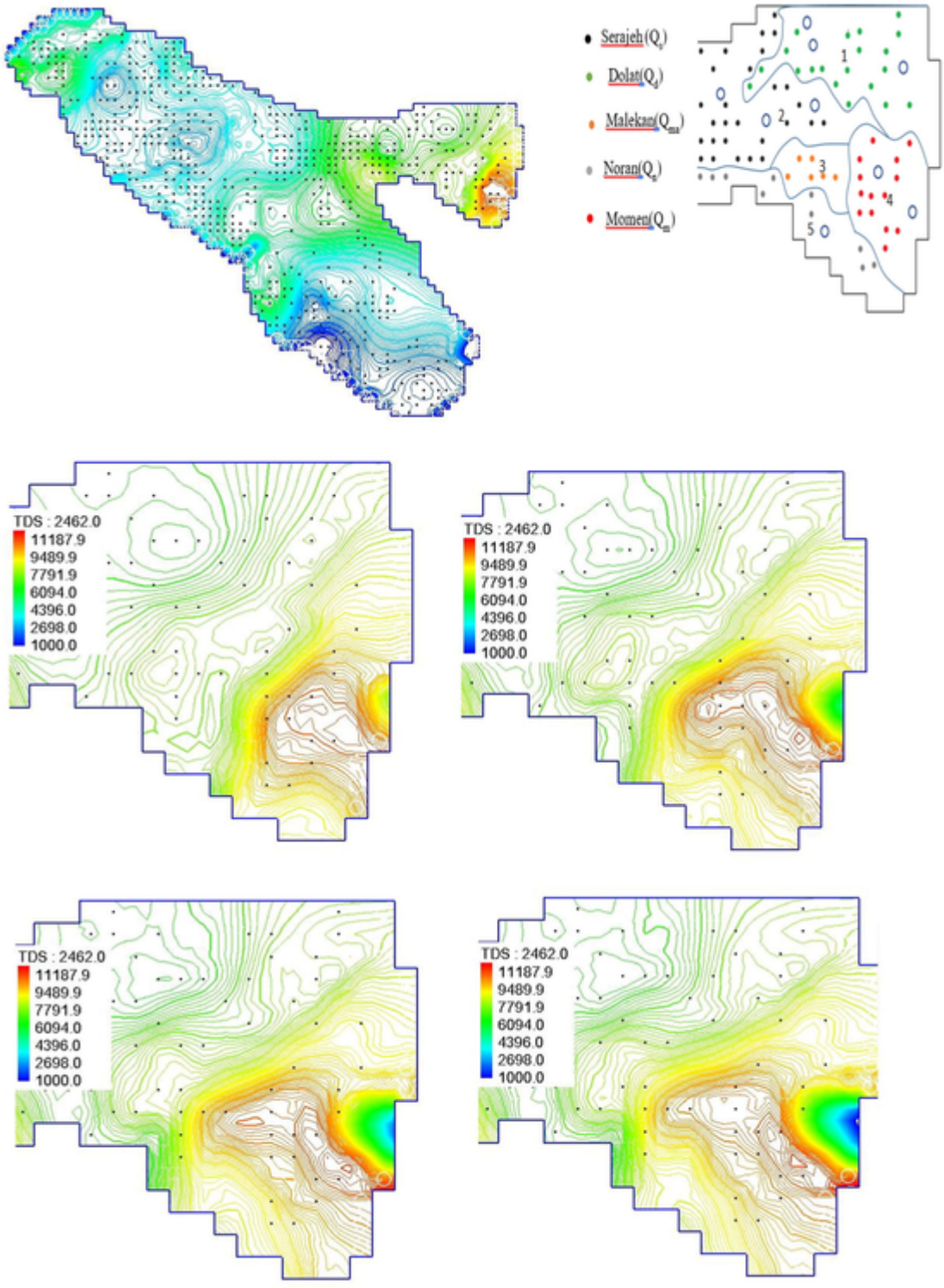


Fig. 13. The contours of TDS concentration corresponding to different values of Nash product (F values) for the five agricultural zones.

rate (74 million cubic meters) from Zone 2 and can be selected as the final groundwater withdrawal scenario.

5. Conclusions

In this paper, a new framework was proposed for the management of aquifers threatened by saltwater intrusion. A conjunctive simulation model was developed using the results of the three models of M5, ANN, and RSM. This model was used as a surrogate for the numerical groundwater simulation model (i.e. SEAWAT), which is a time consuming model when is coupled with an optimization model with a large number of decision variables. To improve the performance of conjunctive surrogate model (CM), the aquifer area was divided into different zones using K-Means clustering technique. To obtain optimal scenarios for groundwater withdrawal, the CV was coupled with an optimization model based on NSGA-II, namely, CV-NSGA-II. The optimization problem had two conflicting objectives of maximizing pumping rates and minimizing SI. Next, the Nash bargaining theory was applied to obtain a socially

Table 8
The main characteristics of the four optimal scenarios and the corresponding values of Nash product.

Scenario	Nash value (F)	$\overline{SI}$ (m)	P^*_1 (mm^3)	P_2 (mm^3)	P_3 (mm^3)	P_4 (mm^3)	P_5 (mm^3)
1	0.0110	5300	416	480	83	97	125
2	0.0124	6400	422	520	95	74	120
3	0.0128	7500	458	838	77	115	132
4	0.0121	7800	508	1306	25	128	15

* Agricultural profit related to cluster i (P_i)

optimal scenrario out of the non-dominated solutions provided by the CV-NSGA-II. The analysis of performance criteria showed that among the three single surrogate models, M5 and ANN model had the best performance in the training phase. The results also showed that for large values of $\overline{SI}$, ANN and M5 models tended to underestimate and overestimate $\overline{SI}$, respectively. This flaw was eliminated using CM-NSGAII, which uses a combination of the results of the three surrogate models. There was a small discrepancy between the Pareto-optimal curve generated by CM-NSGA-II and that of the original SEAWAT-NSGA-II. Also, the proposed CM-NSGA-II model led to 95% reduction in runtime of the simulation–optimization process. The results of applying the bargaining theory to the pareto-optimal scenarios obtained using CM-NSGA-II showed that the most desirable scenario had a Nash value of 0.0128. In addition, scenarios with the Nash value of 0.1124 had maximum pumping rate from wells located far from the shoreline. Therefore, groundwater withdrawal based on this scenario can also increase the quality of extracted water for agricultural uses.

In this paper, only a linear combination of the three surrogate models was considered. Future studies can assess the non-linear combinations of the surrogate models. Also, selecting robust management scenarios out of the Pareto-optimal solutions incorporating the existing uncertainties in aquifer parameters has a potential for future works.

Uncited references

Abd-Elhamid and Javadi (2011), Akaike (1998), Kitanidis (2017), Langevin et al. (2008), Lu et al. (2012), Lu and Luo (2014), Rolle and Kitanidis (2014), Sreekanth and Datta (2014).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jher.2019.11.005.

References

Abd-Elhamid, H.F., Javadi, A.A., 2011. A cost-effective method to control seawater intrusion in coastal aquifers. Water Resour. Manage. 25 (11), 2755–2780.

Alagha, J.S., Seyam, M., Said, M.A.M., Mogheir, Y., 2017. Integrating an artificial intelligence approach with k-means clustering to model groundwater salinity: the case of Gaza coastal aquifer (Palestine). Hydrogeol. J. 1–15.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer, New York, pp. 199–213.

Ay, M., Kisi, O., 2014. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J. Hydrol. 511, 279–289.

Ayoubloo, M.K., Etemad-Shahidi, A., Mahjoobi, J., 2010. Evaluation of wave scour around a circular pile using data mining approaches. Appl. Ocean Res. 32 (1), 34–39.

Banerjee, P., Singh, V.S., Chatttopadhyay, K., Chandra, P.C., Singh, B., 2011. Artificial neural network model as a potential alternative for groundwater salinity forecasting. J. Hydrol. 398 (3), 212–220.

Bhattacharya, B., Price, R.K., Solomatine, D.P., 2007. Machine learning approach to modeling sediment transport. J. Hydraul. Eng. 133 (4), 440–450.

Bhattacharjya, R.K., Datta, B., 2005. Optimal management of coastal aquifers using linked simulation optimization approach. Water Resour. Manage. 19 (3), 295–320.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Chang, H., Liao, Q., Zhang, D., 2017. Surrogate model based iterative ensemble smoother for subsurface flow data assimilation. Adv. Water Resour. 100, 96–108.

Christelis, V., Mantoglou, A., 2016. Coastal aquifer management based on the joint use of density-dependent and sharp interface models. Water Resour. Manage. 30 (2), 861–876.

Christelis, V., Mantoglou, A., 2016. Pumping optimization of coastal aquifers assisted by adaptive metamodelling methods and radial basis functions. Water Resour. Manage. 30 (15), 5845–5859.

Dausman, A.M., Langevin, C., Bakker, M., Schaars, F., 2010. A comparison between SWI and SEAWAT-The importance of dispersion, inversion and vertical anisotropy. The 21st Salt Water Intrusion Meeting. Azores, Portugal 21, 271–274.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.A.M.T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6 (2), 182–197.

Dhar, A., Datta, B., 2009. Saltwater intrusion management of coastal aquifers. I: Linked simulation-optimization. J. Hydrol. Eng. 14 (12), 1263–1272.

Forrester, A.I., Keane, A.J., 2009. Recent advances in surrogate-based optimization. Prog. Aerosp. Sci. 45 (1), 50–79.

Gaur, S., Chahar, B.R., Graillot, D., 2011. Analytic elements method and particle swarm optimization based simulation–optimization model for groundwater management. J. Hydrol. 402 (3), 217–227.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Overview of supervised learning. In: The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, NY, pp. 9–41.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Hosseini, M., Kerachian, R., 2017. A data fusion-based methodology for optimal redesign of groundwater monitoring networks. J. Hydrol. 552, 267–282.

Hussain, M.S., Javadi, A.A., Ahangar-Asr, A., Farmani, R., 2015. A surrogate model for simulation–optimization of aquifer systems subjected to seawater intrusion. J. Hydrol. 523, 542–554.

Kitanidis, P.K., 2017. Teaching and communicating dispersion in hydrogeology, with emphasis on the applicability of the Fickian model. Adv. Water Resour. 106, 11–23.

Ketabchi, H., Ataie-Ashtiani, B., 2015. Coastal groundwater optimization—advances, challenges, and practical solutions. Hydrogeol. J. 23 (6), 1129–1154.

Kourakos, G., Mantoglou, A., 2009. Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models. Adv. Water Resour. 32 (4), 507–521.

Langevin, C.D., Thorne, D.T., Jr, Dausman, A.M., Sukop, M.C., Guo, W., 2008. SEAWAT Version 4: a computer program for simulation of multi-species solute and heat transport, 6–A22. Geological Survey (US).

Lu, C., Chen, Y., Luo, J., 2012. Boundary condition effects on maximum groundwater withdrawal in coastal aquifers. Groundwater 50 (3), 386–393.

Lu, C., Luo, J., 2014. Groundwater pumping in head-controlled coastal systems: the role of lateral boundaries in quantifying the interface toe location and maximum pumping rate. J. Hydrol. 512, 147–156.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematical Statistics and Probability 1 (14), 281–297.

Mahjouri, N., Kerachian, R., 2011. Revising river water quality monitoring networks using discrete entropy theory: the Jajrood River experience. Environ. Monit. Assess. 175 (1–4), 291–302.

Mantoglou, A., Papantoniou, M., 2008. Optimal design of pumping networks in coastal aquifers using sharp interface models. J. Hydrol. 361 (1), 52–63.

Masoumi, F., Kerachian, R., 2008. Assessment of the groundwater salinity monitoring network of the Tehran region: application of the discrete entropy theory. Water Sci. Technol. 58 (4), 765–771.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.

Miraki, S., Zanganeh, S.H., Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Pham, B.T., 2019. Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. Water Resour. Manage. 33 (1), 281–302.

Nikoo, M.R., Beiglou, P.H.B., Mahjouri, N., 2016. Optimizing multiple-pollutant waste load allocation in rivers: an interval parameter game theoretic model. Water Resour. Manage. 30 (12), 4201–4220.

Ouyang, Q., Lu, W., Lin, J., Deng, W., Cheng, W., 2017. Conservative strategy-based ensemble surrogate model for optimal groundwater remediation design at DNAPLs-contaminated sites. J. Contam. Hydrol. 203, 1–8.

Ranjbar, A., Mahjouri, N., 2018. Development of an efficient surrogate model based on aquifer dimensions to prevent seawater intrusion in anisotropic coastal aquifers, case study: the Qom aquifer in Iran. Environ. Earth Sci. 77 (11), 418.

Ranjbar, A., Mahjouri, N., 2019. Multi-objective freshwater management in coastal aquifers under uncertainty in hydraulic parameters. Natural Resour. Res. In Press.

Rao, S.V.N., Thandaveswara, B.S., Bhallamudi, S.M., Srinivasulu, V., 2003. Optimal groundwater management in deltaic regions using simulated annealing and neural networks. Water Resour. Manage. 17 (6), 409–428.

Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48 (7).

Rolle, M., Kitanidis, P.K., 2014. Effects of compound-specific dilution on transient transport and solute breakthrough: a pore-scale analysis. Adv. Water Resour. 71, 186–199.

Roy, D.K., Datta, B., 2017. Fuzzy C-Mean Clustering Based Inference System for Saltwater Intrusion Processes Prediction in Coastal Aquifers. Water Resour. Manage. 31 (1), 355–376.

Roy, D.K., Datta, B., 2017. Multivariate Adaptive Regression Spline Ensembles for Management of Multilayered Coastal Aquifers. J. Hydrol. Eng. 22 (9), 4017031.

Sreekanth, J., Datta, B., 2015. Simulation-optimization models for the management and monitoring of coastal aquifers. Hydrogeol. J. 23 (6), 1155–1166.

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), 245–256.

Sreekanth, J., Datta, B., 2014. Stochastic and robust multi-objective optimal management of pumping from coastal aquifers under parameter uncertainty. Water Resour. Manage. 28 (7), 2005–2019.

UNCORRECTED PROOF

Werner, A.D., Bakker, M., Post, V.E., Vandenbohede, A., Lu, C., Ataie-Ashtiani, B., Barry, D.A., 2013. Seawater intrusion processes, investigation and management: recent advances and future challenges. Adv. Water Resour. 51, 3–26.
Witten, I.H., Frank, E., Hall, M., 2006. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc..
Yasa, R., Etemad-Shahidi, A., 2014. Classification and regression trees approach for predicting current-induced scour depth under pipelines. J. Offshore Mech. Arct. Eng. 136 (1), 011702.

UNCORRECTED PROOF

View publication stats