Short-term prediction of groundwater level using improved random forest regression with a combination of random features

Xuanhui Wang^{1,2,3} · Tailian Liu¹ · Xilai Zheng^{2,3} · Hui Peng^{2,3} · Jia Xin^{2,3} · Bo Zhang^{2,3}

Received: 7 March 2017 / Accepted: 7 June 2018 / Published online: 24 July 2018
© The Author(s) 2018

Abstract

To solve the problem where by the available on-site input data are too scarce to predict the level of groundwater, this paper proposes an algorithm to make this prediction called the canonical correlation forest algorithm with a combination of random features. To assess the effectiveness of the proposed algorithm, groundwater levels and meteorological data for the Daguhe River groundwater source field, in Qingdao, China, were used. First, the results of a comparison among three regressors showed that the proposed algorithm is superior in terms of forecasting variations in groundwater level. Second, the results of experiments were used to show the comparative superiority of the proposed method in terms of training time and complexity of parameter optimization. Third, using the proposed algorithm, the highest prediction accuracy was achieved by employing precipitation $P(t-2)$, temperature $T(t)$, and groundwater level $H(t)$ as the best time lag. This improved random forest regression model yielded higher accuracy in forecasting the variation in groundwater level. The proposed algorithm can also be applied to cases involving low-dimensional data.

Keywords Groundwater level prediction · Improved random forest regression · Least squares support vector regression · Daguhe River groundwater source field

Abbreviations

ANN	Artificial neural network
CCA	Canonical correlation analysis
CCF-CRF	Canonical correlation forests with a combination of random features
Forest-RC	Features combination based random forest
GWL	Groundwater level
LS-SVR	Least squares support vector regression
MAE	Mean absolute error
R	Coefficient of correlation
RF	Random forest
RFR	Random forest regression

RMSE	Root-mean-squared error
SVM	Support vector machine

Introduction

Fluctuations in groundwater level (GWL) can be used to evaluate groundwater stability and flow as well as the characteristics of the aquifer. Moreover, groundwater is the main source of farmland irrigation in major agricultural regions. Agricultural water managers assess whether groundwater is needed during sowing season and how much is likely to be provided (Barlow and Clark 2011; Dyer et al. 2015; Kebede et al. 2014). Although most government agencies collect GWL data once or twice a year in major agricultural areas, this is not sufficient for short-term studies. Therefore, it is necessary to achieve acceptable prediction accuracy of GWL variations when previous information is not available and the computational sources are limited. It is well known that the dynamic variation in GWL is influenced by meteorological phenomena, urbanization, tidal effects, and land subsidence (Khaki et al. 2015). Of these, such meteorological parameters as atmospheric pressure, frost, precipitation,

✉ Xilai Zheng
zhxilai@ouc.edu.cn

¹ Science and Information College, Qingdao Agricultural University, Qingdao 266109, China

² Key Lab of Marine Environmental Science and Ecology, Ministry of Education, Ocean University of China, Qingdao 266100, China

³ Shandong Provincial Key Laboratory of Marine Environment and Geological Engineering, Ocean University of China, Qingdao 266100, China

and evapotranspiration are considered important indicators of fluctuations in GWL.

Models to simulate groundwater can be classed into two main groups: physical models and data-driven models. For physical models, an appropriate synthesis of the parameters of the aquifer is used to determine the spatial variation in underground space. This information is challenging to obtain even through expensive on-site surveys, which further increase computational cost. Therefore, the physical domain needs to be partitioned to obtain a numerical solution (Taormina et al. 2012). An appropriate alternative to the physical model is the data-driven model, which can provide accurate predictions without a costly calibration time when data are insufficient and the physical mechanism is not the focus of research (Mohanty et al. 2015). Artificial neural network (ANN)-based techniques and the support vector machine (SVM) are data-driven models treated as standard nonlinear estimators and can overcome the limitations of the physical model (Behzad et al. 2010).

In groundwater hydrology, ANNs have been used in such applications as groundwater level prediction (Emamgholizadeh et al. 2014; He et al. 2014; Khaki et al. 2015). However, it is known that ANN models incur the disadvantages of local minima and problems of overfitting. The SVM can overcome these drawbacks (Hosseini and Mahjouri 2016). Comparative studies on the SVM and the ANN have been performed in the context of GWL prediction. Behzad et al. (2010) compared the SVM and ANN for predicting transient GWL under variable pumping and weather conditions. Yoon et al. (2011) developed and compared two time-series forecasting models using ANN and SVM and applied them to forecast GWL fluctuations in a coastal aquifer recharged from precipitation and tidal effects. Yoon et al. (2016) compared the recursive prediction performance of the SVM and ANN for the long-term prediction of GWL and concluded that SVM is a better substitute for ANN in terms of accuracy and robustness for the prediction of GWL fluctuation.

Although SVM is superior to ANN in reflecting the dynamic variation in GWL, it takes more time because of trial and error (Raghavendra and Deka 2014). It is also sensitive to outliers and redundant data (Suykens et al. 2002). On the contrary, results of the prediction of random forest (RF) are unaffected by outliers and redundant data (Rodriguez-Galiano et al. 2014). Comparative studies on RF and SVM have been carried out from different perspectives. For example, Pal (2005) compared RF with SVM in terms of classification accuracy, training time, and user-defined parameters. They were also compared on four criteria in mineral prospectivity modeling (Rodriguez-Galiano et al. 2015). In these case studies, the authors concluded that the RF model exhibited stronger predictive capabilities and higher robustness, with a lower complexity of parameterization and a shorter training time for optimization than the SVM.

RF has been widely used in surface and ground water hydrology. Many studies have applied RF to study groundwater from different aspects, such as its vulnerability to nitrate pollution (Rodriguez-Galiano et al. 2014), the dissolution of organic nitrogen in it (Wang et al. 2016), and the potential mapping of groundwater (Rahmati et al. 2016). Moreover, the RF model has been compared with the SVM in terms of surface water level prediction (Li et al. 2015). However, no study to date has applied RF to the prediction of GWL owing to low-dimensional input features.

Breiman (2001) introduced a random forest algorithm (Forest-RC) that used linear combinations of input variables when dealing with low-dimensional input data. Although Breiman (2001), Hamza and Larocque (2005) and Moudani (2013) have shown that Forest-RC yields better results than those obtained with no linear combination, there are five key limitations to this approach:

1. The extended number of dimensions in Forest-RC is limited when the original data contain few dimensions of feature.
2. The original feature information, which is important for regression and has a certain impact on its results, is removed by Forest-RC.
3. The information in the new feature varies with the value of L , which is the number of selected variables that are combined. Furthermore, it directly affects the construction of the decision trees. Experiments have shown that different values of L lead to different generalization capabilities (Luo et al. 2016).
4. A fixed value of L is contrary to randomness, which is introduced by choosing training samples randomly and using a random subset of all features as the set of splitting rules. The value of L must be adjusted for different prediction horizons for GWL, which limits the versatility of the Forest-RC algorithm in practical applications.
5. Forest-RC uses orthogonal (i.e., axis-aligned) decision trees that only recognize axis-parallel splits of the feature space and lead to poorer performance of trees than oblique decision trees. Zhang and Suganthan (2015) showed that oblique random forests achieved better performance than their conventional axis-parallel counterpart in computation time and classification accuracy.

Based on the above issues, in this paper, we propose a method of GWL modeling based on canonical correlation forests with a combination of random features (CCF-CRF) that expands the low-dimensional feature vector space to high-dimensional space and uses canonical correlation components for oblique splits. A comparison of performance among CCF-CRF, random forest regression (RFR) algorithm, and least squares support vector regression (LS-SVR) was conducted and showed that CCF-CRF can provide the most accurate

short-term GWL forecasts. For this purpose, we chose the monitoring well Guxian along the groundwater source for the Daguhe River for a case study.

Materials and methods

Random forest regression algorithm (RFR)

Based on classification and regression trees (CART), random forest (RF) builds a forest from bootstrap samples of observed data. Having generated multiple decision trees in the process of training, RF is designed to generate an output by majority vote (for classification) and the average of the single-tree method (for regression) (Breiman 2001). GWL as the output of this prediction is a continuous variable. Thus, we focus on the regression model of RF.

A regression tree (RT), where each non-leaf node contains a set of decision rules and each leaf node is the outcome of a prediction, is a form of decision tree (DT) (Quinlan 1993; Rodriguez-Galiano et al. 2014).

The two user-defined parameters of RF are B , the number of trees in the forest, and D , the number of features used to split the nodes. The default value of B is 500, but is reset according to application. One third of the variables for the regression task are set to the default value of D .

Steps of the RF regression algorithm are as follows (for full details, see Breiman 2001):

1. Different bootstrap samples X_i (i = bootstrap iteration) are randomly drawn from the original dataset X . Two-thirds of the samples are included in a bootstrap sample and one third as the out-of-bag samples. Each tree is constructed to correspond to a particular subset of the bootstrap.
2. At a node in each tree, a new split is randomly selected from all indices, and the input variable with the lowest mean square error (MSE) is chosen as the splitting criterion of the regression tree.
3. The data splitting process in each internal node is repeated according to the above steps until all randomized trees have been grown and a stop condition is reached.
4. The final results of regression can be calculated as follows, where B stands for the number of trees in the forest and T_b represents each tree:

$$\hat{y}(x_i) = \frac{1}{B} \sum_{b=1}^B T_b(x_i). \quad (1)$$

Least squares support vector regression (LS-SVR)

An extension of the standard support vector regression (SVR) (Smola et al. 2004; Vapnik 1995) is least squares

support vector regression (LS-SVR) (Pelckmans et al. 2002; Suykens and Vandewalle 2000). LS-SVR is a series of deformations that can replace the quadratic programming problem with the linear equation problem. The calculation can thus be made easier, and the time needed to obtain a solution can be effectively reduced.

The basic principle of LS-SVR is introduced in this section (Suykens and Vandewalle 2000). The training sample set was assumed to be $\{X_i, Y_i\}_{i=1}^N$, where $X_i \in R^k$ are the input data and $Y_i \in R$ the output data. The relationship between the input and output data is as follows:

$$Y(X) = \sum_{i=1}^N \alpha_i G(X, X_i) + b \quad (2)$$

where G is a kernel-based function and α_i and b are parameters of the regression. The primary method of LS-SVR involves converting the parameter estimation problem into the following quadratic programming problem:

$$\min_{\alpha, b, e} \frac{1}{2} \alpha^T \alpha + \frac{\gamma}{2} e^T e \quad (3)$$

$$\text{s.t. } Y_i = \sum_j \alpha_j G(X_i, X_j) + b + e_i, \quad i = 1, \dots, N \quad (4)$$

$\gamma \in R^+$ is a penalty factor, $e_i \in R$ is the bias term, G is the radial basis function (RBF), a kernel function that is the best choice for researchers because it has higher accuracy and reliability than other kernel functions (Raghavendra and Deka 2014), and α_i and b are parameters of the regression. With $\hat{Y} = (Y_1, Y_2, \dots, Y_N)^T \in R^N$ and $1_N = (1, 1, \dots, 1)^T \in R^N$, the solutions are given by

$$\begin{bmatrix} \Omega + \frac{I_N}{\gamma} & 1_N \\ 1_N^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} \hat{Y} \\ 0 \end{bmatrix} \quad (5)$$

where $\Omega_{ij} = G(X_i, X_j)$ is the kernel matrix and I_N the identity matrix of order N .

Canonical correlation forests with a combination of random features (CCF-CRF)

The algorithm improves the predictive performance in two respects:

(1) the use of a combination of random features that introduces more randomness to build the decision tree, and (2) the use of canonical correlation analysis (CCA) to generate candidate hyperplane splits and projection bootstrapping to construct ensembles of the oblique decision tree.

1. A combination of random features

Two ingredients involved in the generalization error for random forests: the strength of a predictor in the forest and the correlation between trees (Breiman 2001). When the strength is fixed, the lower the correlation coefficient, the smaller the generalization error. Luo et al. (2016) proposed that one solution that reduces the correlation coefficient involves expanding the low-dimensional feature vector space to a high-dimensional space.

The original RF method can handle cases with only a few inputs. The only viable solution to this problem is to increase the dimensionality of the feature space, which can be done by creating linear combinations of input features. Forest-RC (Breiman 2001) is one such method and was introduced to define more features by using random linear combinations of input variables.

In Forest-RC, N_L features are selected randomly from the original features at each node and linearly combined together with a random number uniformly distributed in the interval $[-1, 1]$. We thus get the variable names V , whose formulation is as follows:

$$V = \sum_{i=1}^L K_i V_i, \quad K_i \in [-1, 1] \quad (6)$$

Thus, more features are defined by taking random linear combinations of the input variables, and the nodes are split using the optimal predictive variable among these new features.

As mentioned in the “Introduction” section, Forest-RC has five defects. To overcome them and improve the predictive performance of the ensemble, this paper proposes an algorithm called canonical correlation forests with a combination of random features (CCF-CRF). The value of L induces randomness to enable the feature space to contain a greater number of novel and different features. The original m -dimensional feature space is extended to n -dimensions, and this relationship is described by the following equation:

$$n = \sum_{i=1}^m C_M^i. \quad (7)$$

It is clear from Eq. (7) that the new n -dimensional feature space not only contains the original feature space, but also contains the linear combination of arbitrary features. CCF-CRF is thus an approach that creates a higher-dimensional feature space than that of Forest-RC and RFR. The processing of the CCF-CRF algorithm is described in detail in Fig. 1.

CCF-CRF, where the split at each node is based on completely random linear combinations of features instead of a single one, has a lower probability of choosing the

same attribute for each node, which reduces the similarity of the decision trees. To reducing the correlation coefficient and the generalization error, more randomness is introduced into the building of individual regression trees by the CCF-CRF algorithm.

2. CCA and projection bootstrapping

Canonical correlation analysis (CCA) (Hotelling 1936) is used to calculate pairs of linear projections to maximize the correlation coefficient between matrices in a co-projected space. In the first step, features λ appearing in Algorithm 1 are sampled without replacement from the input feature set F_s , which is generated from the process in part 2.2 of this algorithm. In the next step, a new training set $\{X', Y'\}$ is selected from $\{X_{(:,\lambda)}, Y\}$ using the bootstrap sampling technique. CCA coefficients relating the features to the outputs are then calculated based on the new training set. The canonical coefficients Φ corresponding to X are used to generate the new features, which are obtained by projecting the original features into the canonical component space. Furthermore, the function FINDBESTSPLIT obtains the best split ξ by calculating the maximum information gain (Quinlan 1986).

From the above analysis, it is clear that the main difference between RF and CCF-CRF is that the latter uses CCA to analyze the relationship between the output and the features and then splits the nodes using an exhaustive search in the projected space.

Study area and data processing

Study site and data description

The study site was located at the groundwater source field of Daguhe River in Qingdao, China. The monitoring well Guxian, located in the northwest of the study area, was used as an example to verify the effectiveness of the CCF-CRF model. The groundwater source field of Daguhe River is among the important public drinking water sources in Qingdao. It is located in the northwest of the city and covers an area of approximately 4600 km², with the temperature ranging from -3.3 °C in January to 25.3 °C in July and average annual precipitation of 625.3 mm. The geographical position of the site is shown in Fig. 2.

Figure 3 shows a typical geological profile along section $a-a'$ nearby Guxian. The major aquifer systems of the study site were constituted by quaternary alluvium and diluvium, consisting of unconsolidated gravel, sand, silt, and clay with a width of 5–7 km. The maximum width was more than 10 km, and the thickness ranged from 10 to 20 km. The aquifer system was discretized vertically into two layers. There

Algorithm: canonical correlation forests with a combination of random feature algorithm (CCF_CRFC)

Input: initial training sample set T

the number of input features to sub-sample $F \in \{1, \dots, D\}$

1、Let N to be the number of training samples in set T , and let F to be the number of input features of each training sample in set T .

2、Construct the regression tree (RT) by the training set T :

2.1 Using the bootstrap method, taking N samples randomly from T as a training sample set to building the i^{th} regression tree method by sampling with replacement.

2.2 Building the i^{th} regression decision tree on N . The processing is:

A. For $j = 1$ to $F_s = \lfloor \text{rand}(1, F) \rfloor =$

- a) Generate a random number N_i which values is in $[1, F]$
- b) select N_i input features from the F initial input features
- c) For each of the N_i input features, generate a uniform random number between $[-1, +1]$ as weight vector.
- d) At a given node, N_i variables are added together with coefficients that are uniform random numbers on $[-1, 1]$, which was made as j^{th} candidate splitting feature of the node.

B. $\{X, Y\}$: training samples at the i^{th} regression decision tree

- a) Randomly selecting λ features by sampling from $\{1, \dots, F_s\}$ without replacement.
- b) $\{X', Y'\} \leftarrow$ bootstrap sampling from $\{X_{(c,\lambda)}, Y\}$
- c) $[\Phi, \cdot] = \text{CCA}(X', Y')$
- d) $U \leftarrow X\Phi$
- e) $[\xi, \text{gain}] = \text{FINDBESTSPLIT}(U)$

C. Use ξ candidate splitting features calculated in step B to split the node, then select the candidate feature which has the lowest MSE to be the splitting features of the node.

D. Recursive call the step A~C respectively to construct each branch until a stop condition previously specified is reached.

2.3 Repeat step 2.1 and 2.2, until the establishment of k regression trees which constitute the random forest mode.

3、Evaluate the performance of the random forests by the OOB data.

4、Use the random forests to perform the regression.

Illustration:

Canonical correlation analysis (CCA) (Hotelling, 1936) is a deterministic method for calculating pairs of linear projections that maximize the correlation between co-projected matrices. Decision Trees (DT) is trained by using CCA to find feature projections and then selecting the best splits in this projected space in order to improve the prediction accuracy.

Fig. 1 Description of CCF-CRF algorithm

were particles of small size in layer 1, which was covered with silty clay, and their width ranged from 2 to 5 km. Layer 2 was stocked with sand and gravel, with larger particle sizes with a thickness ranging from 4 to 8 km. The bottom could be regarded as aquiclude composed of clay rock and sandstone and is also known as the lower confining bed.

In a majority of the area, the rainy season peaked in May through July. The most obvious characteristics of this

area were the vegetation cover and the absence of a significant variation in topography. The depth to the groundwater was small, with a thin silt-like clay layer in this area, which created conditions favorable for rainfall infiltration. Precipitation was thus the primary and most critical variable affecting short-term fluctuations in groundwater level within the study site.

Fig. 2 Map of situation of the studied site (Guxian Station)

Fig. 3 Simple geological profile along the section $a-a'$ near by Guxian well

The precipitation in millimeters and temperature data in Celsius were obtained from a meteorological station located near the Guxian GWL station. Daily data for the groundwater level for the site were collected from a remote monitoring system designed to provide accurate scientific basis for the monitoring, exploitation, and protection of groundwater from the Daguhe River groundwater source field. Daily

GWL values used in this modeling were the averages of two measurements per day, whereas the corresponding daily precipitation value was the total precipitation measured over a 24-h period, and the corresponding daily minimum temperature was the minimum value of all observations within a day.

From Fig. 4, it is clear that the lowest GWL (during the study period) obtained in July 2013 and 2014, with warmer

months yielding the lowest GWL. This can be attributed to higher evaporation rates owing to higher temperatures in the summer, which motivated higher groundwater extractions and soil-moisture deficits. On the contrary, highest GWL occurred in spring under natural conditions because recharge was relatively high and groundwater pumping extraction remained relatively low.

Data preprocessing and performance criteria

The RFR, LS-SVR, and CCF-CRF models for GWL forecasting at the Guxian well station were written in MATLAB. The applied data were gathered from January 1, 2013, to November 19, 2014, at the Guxian well. For each dataset, 66.5% of the data were selected as the initial training samples set T , and the remaining 33.5% were used as test samples. To calibrate the parameters of the LS-SVR model, the training samples set were divided into two parts: a training set and a calibration set. In the former, system behavior had been learned and the correlated patterns between the input dataset and the target values identified by LS-SVR. In the latter, the optimal model parameters were determined under the rule of minimizing error through a trial-and-error process. In the testing stage, the performance of a fully specified predictor was assessed on criteria for model performance. However, a separate calibration set was not needed for the RF to obtain an unbiased estimate of the test set error because out-of-bag estimation is the internal cross-validation of the training set and a good tool for optimizing the parameters.

Before the data were treated by these three models, all input and output data in the training process were normalized using the mean (X_{mean}) and maximum (X_{max}) values, as described

in Eq. (8), so that the variables (X) in the training and testing datasets ranged from -1 to 1 . In Eq. (8), X_{norm} , X , X_{mean} , and X_{max} represent the normalized value, the real value, the mean value, and the maximum value, respectively:

$$X_{\text{norm}} = \frac{X - X_{\text{mean}}}{2X_{\text{max}}} \quad (8)$$

The root-mean-square error (RMSE), the mean absolute error (MAE), and the correlation coefficient (R) are usually employed as evaluation metrics of the model used in this study. The RMSE is a standard metric for model error and represents the degree of linear relationship between the observed GWL values and the forecasted values. As the errors are squared before they are averaged, it is very sensitive to large errors in a set of measurement data. The MAE is another useful method that uses the deviation between the forecasted values and the actual values to reflect the accuracy of the system. Equation (11) represents the MAE. The predictive capability of the model was evaluated by RMSE and MAE from different points of view. Our analysis indicates that MAE is a good measure of overall error in the training and testing sets, and RMSE measures the goodness of fit relevant to high values. The best fit between the observed and estimated values would have been $R=1$, $\text{RMSE}=0$, and $\text{MAE}=0$. These parameters were calculated by using the following equations:

$$R = \frac{\sum_{i=1}^n (Q_i^o - \bar{Q}^o)(Q_i^p - \bar{Q}^p)}{\sqrt{\sum_{i=1}^n (Q_i^o - \bar{Q}^o)^2 \sum_{i=1}^n (Q_i^p - \bar{Q}^p)^2}} \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_i^p - Q_i^o)^2} \quad (10)$$

Fig. 4 Time-series plot collected at Guxian studied site

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Q_i^p - Q_i^o| \quad (11)$$

where n is the number of input samples, Q_i^o and Q_i^p are the observed and predicted GWLs at the i th time step, respectively, $\bar{Q}^o$ and $\bar{Q}^p$ represent the mean values of the observed and predicted GWLs, respectively.

Results and discussion

Comparison of RFR, LS-SVR, and CCF-CRF to determine the best model

The performance of the original RFR, LS-SVR, and CCF-CRF models is compared in Fig. 5 for forecasting variations in 1-, 3-, 5-, 7-, and 10-day groundwater levels at the Guxian site. All performance evaluation measures of the CCF-CRF model, including R , RMSE, and MAE, are shown in red. The same input structures and prediction horizons as in the CCF-CRF model were introduced to the LS-SVR model, whereas the RFR model had three driving factors: temperature, precipitation, and previous GWLs. The number of decision trees (B) was set to 100 because the differences in computation time were more obvious, but those in accuracy were small when the number of decision trees was more

than 100 (Rodriguez-Galiano et al. 2012). The default value of parameter D was 1/3 of the total number of variables.

According to Fig. 5, the LS-SVR model yielded the best R , RMSE, and MAE scores in 1-day prediction. However, as the lag time of the forest increased, CCF-CRF gradually began to exhibit performance superior to that of the RFR and LS-SVR models in terms of overall trend. Furthermore, Fig. 5 shows that RFR outperformed LS-SVR for long-term predictions primarily because it split each node using only the best feature from a random subset of features. This operation can be seen as internal feature selection with no equivalent in the LS-SVM. Above all, it can be concluded that the CCF-CRF model was the most efficient for daily groundwater forecasting for two reasons. First, the split at each node was based on completely random linear combinations of features instead of a single one and had a lower probability of choosing the same attribute for each node, which reduced similarity among the decision trees. Second, as an alternative to the bagging technique (Breiman 1996), the projection bootstrap technique was used to improve accuracy and diversity within decision trees. It was found that 1-day prediction of the three models obtained better performance than 3-, 5-, 7-, and 10-day predictions. The accuracy of the forecasted results degraded as the lead time of the forecast increased. These results are consistent with those of previous studies (Chang et al. 2015).

Another important criterion for model comparison is computation time. In this context, Table 1 lists the training times and complexity for parameter optimization for

Fig. 5 The performance comparison of the original RFR model, LS-SVR model, and CCF-CRF model for forecasting the 1-, 3-, 5-, 7-, and 10-day-ahead groundwater level variations at Guxian site

both regressors. The first row of Table 1 shows the optimal parameter values corresponding to each model. For RF, the number of trees (B) was set to 100 and the number of features, randomly selected at each node (F), was set to three. For CCF-CRF, the number of trees (B) was set to 100 and the number of features (F) to seven. The training times under the optimal parameter configuration were given in M . The time required to perform the parameter optimization is compared in N .

Having obtained the optimal parameters, CCF-CRF was found to be 17 times faster than LS-SVR. On 330 training samples, parameter optimization for LS-SVR took nearly 36 times longer than that for CCF-CRF. Pal (2005) reached the same conclusions, whereby LS-SVR was known for having a longer training time than RF because it often requires an optimization phase that is seldom straightforward, but CCF-CRF needs only slight parameter tuning for a short training time for optimization. This is shown in Table 1, where the computation time of the RFR ensemble was small, but Fig. 5 shows that the proposed CCF-CRF was more efficient than RFR in most cases, particularly for long-term predictions. For CCF-CRF, there was a trade-off between performance and computation time.

Table 1 Training times in seconds for LS-SVR, RFR, and CCF-CRF

	LS-SVR	RFR	CCF-CRF
Optimized parameter configuration	$\gamma = 10^8$ $e = 5.5$	$B = 100$ $F = 3$	$B = 100$ $F = 7$
M	298	28	39
N	1399	57	71

γ is a penal factor, and e is the bias terms of LS-SVR

Table 2 Performance of the CCF-CRF model with different lag times during the testing periods for 1-, 3-, 5-, 7-, and 10-day-ahead groundwater level forecasting for Guxian well station

	$H(t+1)$		$H(t+3)$		$H(t+5)$		$H(t+7)$		$H(t+10)$	
	R	RMSE	R	RMSE	R	RMSE	R	MSE	R	MSE
$P(t-1)T(t)H(t)$	0.9410	0.1649	0.9006	0.1942	0.8650	0.2254	0.8176	0.2634	0.7690	0.2955
$P(t-2)T(t)H(t)$	0.9581	0.1338	0.9213	0.1476	0.8704	0.2238	0.8361	0.2478	0.8223	0.2668
$P(t-3)T(t)H(t)$	0.9553	0.1400	0.9063	0.1929	0.8368	0.2449	0.8239	0.2642	0.8089	0.2855
$P(t-4)T(t)H(t)$	0.9482	0.1349	0.9087	0.1876	0.8580	0.2300	0.8197	0.2575	0.7728	0.2893
$P(t)T(t-1)H(t)$	0.9572	0.1375	0.9012	0.1873	0.8552	0.2319	0.8372	0.2619	0.7420	0.3023
$P(t)T(t-2)H(t)$	0.9479	0.1353	0.9051	0.1493	0.8597	0.2394	0.8278	0.2651	0.7906	0.2855
$P(t)T(t-3)H(t)$	0.9509	0.1349	0.9073	0.1900	0.8548	0.2341	0.8006	0.2713	0.7020	0.3200
$P(t)T(t-4)H(t)$	0.9571	0.1375	0.9013	0.1766	0.8643	0.2315	0.8301	0.2621	0.6846	0.3277
$P(t)T(t)H(t-1)$	0.9401	0.1480	0.9004	0.1500	0.8424	0.2419	0.8244	0.2627	0.8015	0.2827
$P(t)T(t)H(t-2)$	0.9127	0.1847	0.8531	0.2330	0.8277	0.2492	0.8011	0.2802	0.7064	0.3111
$P(t)T(t)H(t-3)$	0.9346	0.1700	0.9008	0.1992	0.8298	0.2468	0.8293	0.2642	0.8003	0.2943
$P(t)T(t)H(t-4)$	0.9267	0.1746	0.8976	0.2205	0.8177	0.2615	0.8189	0.2567	0.7947	0.2876

P represents daily precipitation value, T represents daily minimum temperature value, and H represents daily GWL value. The lowest RMSE and highest R are in bold font

Comparison between horizons to determine the optimal time lag

The lag times of precipitation, temperature, and historical GWL have a significant influence on the predicted GWL (Yoon et al. 2011). Therefore, based on the results of a comparison among the results reported in the previous subsection, a series of time lags were selected to predict daily GWL using the CCF-CRF model. From Table 2, it is evident that the model with time lag $P(t-2)T(t)H(t)$ represents the best performance at different horizons during the testing stage, which confirms the retardation effect of GWL fluctuation due to rainfall at this site. This can be attributed to the construction of the aquifer on the site. It was also found that a 1-day yielded obtained better performance than 3-, 5-, 7-, and 10-day predictions. The accuracy of the forecasted results decreased as the lead time of the forecast increased. For a long lag time (7 and 10 days), the CFRC-CCF model with a lag time of $P(t-2)T(t)H(t)$ performed well, which is consistent with the results shown in Fig. 5.

Predicting groundwater level

Once the best model scenario and corresponding time lag were determined, the optimal model was used to predict groundwater levels for 1-, 3-, 5-, 7-, and 10-day forecasts. Based on the best model and the time lag, Fig. 6 shows the scatter plots of optimal models for deferent horizons in the testing period. It is clear that 1-day GWL predictions were less scattered and closer to the straight line than those the other predictions. In general, the CCF-CRF model showed impressive performance with respect to R and RMSE for all horizons. It was also found that it underestimated high GWLs and overestimated low ones in predicting extreme

Fig. 6 Comparison of observed and predicted groundwater level of the optimal model for 1, 3, 5, 7, and 10 days (corresponding to a–e, respectively) ahead during the testing period

occurrences, as found by Wunsch et al. (2018) and Emamgholizadeh et al. (2014). Therefore, GWL prediction should be divided into low and high levels in case of a rapid drop or rise.

It was this verified that the CCF-CRF model with $P(t - 2)T(t)H(t)$ lag time offers relatively good agreement between the observed values and their corresponding measured values.

Conclusions

This paper applied CCF-CRF to short-term GWL prediction at the Guxian well at the groundwater source field of the Daguhe River. To evaluate the effectiveness of CCF-CRF, its accuracy was compared with LS-SVR and RFR, and results show that it can generate a more accurate estimation of GWL for various prediction horizons (1, 3, 5, 7, 10 days). It was also found that the CCF-CRF model offers a better trade-off between prediction performance and computation time than both the other two algorithms. Based on the optimal model, the highest prediction accuracy was achieved using precipitation $P(t - 2)$, temperature $T(t)$, and groundwater level $H(t)$ as time lag. The CCF-CRF yielded excellent performance, particularly over longer prediction horizons with sparser data pairs.

Overall, the results of the case study are favorable and show that CCF-CRF is a promising prediction tool in groundwater hydrology. Moreover, this paper provided case study-based illustrations that show that it is suitable for situations where only low-dimensional input data are available. This study used data from only one station, and more data from other areas can be used to verify the conclusions of this study. Furthermore, research is needed to improve predictive accuracy when there is sudden change in GWL over consecutive time periods.

Acknowledgements This study was supported by the National Key Research Project (2016YFC0402810), the National Natural Science Foundation of China (51409236), and Research on Intelligent Decision Support System of Tea Garden Fertigation Based on Big Data (ZR2017LC027) of the Shandong Natural Science Foundation.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Barlow JRB, Clark BR (2011) Simulation of water-use conservation scenarios for the Mississippi Delta using an existing regional groundwater flow model. USGS
- Behzad M, Asghari K, Coppola EA (2010) Comparative study of SVMs and ANNs in aquifer water level prediction. *J Comput Civ Eng* 24(5):408–413
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Chang J, Wang G, Mao T (2015) Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. *J Hydrol* 529:1211–1220
- Dyer J, Mercer A, Rigby JR, Grimes A (2015) Identification of recharge zones in the lower Mississippi River alluvial aquifer using high-resolution precipitation estimates. *J Hydrol* 531(Part 2):360–369. <https://doi.org/10.1016/j.jhydrol.2015.07.016>
- Emamgholizadeh S, Moslemi K, Karami G (2014) Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resour Manag* 28(15):5433–5446. <https://doi.org/10.1007/s11269-014-0810-0>
- Hamza M, Larocque D (2005) An empirical comparison of ensemble methods based on classification trees. *J Stat Comput Simul* 75(8):629–643. <https://doi.org/10.1080/00949650410001729472>
- He Z, Zhang Y, Guo Q, Zhao X (2014) Comparative study of artificial neural networks and wavelet artificial neural networks for groundwater depth data forecasting with various curve fractal dimensions. *Water Resour Manag* 28(15):5297–5317. <https://doi.org/10.1007/s11269-014-0802-0>
- Hosseini SM, Mahjouri N (2016) Integrating support vector regression and a geomorphologic artificial neural network for daily rainfall-runoff modeling. *Appl Soft Comput J* 38:329–345. <https://doi.org/10.1016/j.asoc.2015.09.049>
- Hotelling H (1936) Relations between two sets of variates. *Biometrika* 28:321–377
- Kebede H, Fisher DK, Sui R, Reddy KN (2014) Irrigation methods and scheduling in the delta region of Mississippi: current status and strategies to improve irrigation efficiency. *Am J Plant Sci* 5:2917
- Khaki M, Yusoff I, Islami N (2015) Simulation of groundwater level through artificial intelligence system. *Environ Earth Sci* 73(12):8357–8367. <https://doi.org/10.1007/s12665-014-3997-8>
- Li B, Yang G, Wan R, Dai X, Zhang Y (2015) Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. *Hydrol Res* 47(S1):69–83. <https://doi.org/10.2166/nh.2016.264>
- Luo Y, Huang D, Liu P (2016) An novel random forests and its application to the classification of mangroves remote sensing image. *Multimed Tools Appl* 75:9707–9722. <https://doi.org/10.1007/s11042-015-2906-9>
- Mohanty S, Jha MK, Raul SK, Panda RK, Sudheer KP (2015) Using artificial neural network approach for simultaneous forecasting of weekly groundwater levels at multiple sites. *Water Resour Manag* 29(15):5521–5532. <https://doi.org/10.1007/s11269-015-1132-6>
- Moudani W (2013) Dynamic features selection for heart disease classification. *World Acad Sci Eng Technol* 7:105–110
- Pal M (2005) Random forest classifier for remote sensing classification. *Int J Remote Sens* 26(1):217–222. <https://doi.org/10.1080/0143160412331269698>
- Pelckmans K, Suykens JAK, Van Gestel T, De Brabanter J, Lukas L, Hamers B, MoorB De, Vandewalle J (2002) LS-SVMlab: a MATLAB/C toolbox for least squares support vector machines. ESAT-SISTA, K.U. Leuven, Leuven

- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1(1):81–106
- Quinlan JR (1993) C4.5 programs for machine learning San Francisco, 1st edn. Morgan Kaufmann Publishers Inc., California, p 303
- Raghavendra NS, Deka PC (2014) Support vector machine applications in the field of hydrology: a review. *Appl Soft Comput* 19:372–386. <https://doi.org/10.1016/j.asoc.2014.02.002>
- Rahmati O, Reza H, Melesse AM (2016) Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. *Catena* 137:360–372. <https://doi.org/10.1016/j.catena.2015.10.010>
- Rodriguez-galiano VF, Ghimire B, Rogan J, Chica-olmo M, Rigol-sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogram Remote Sens* 67:93–104. <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L (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. <https://doi.org/10.1016/j.scitotenv.2014.01.001>
- Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol Rev* 71:804–818. <https://doi.org/10.1016/j.oregeorev.2015.01.001>
- Smola AJ, Sch B, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14(3):199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Suykens JAK, Vandewalle J (2000) Recurrent least squares support vector machines. *IEEE Trans Circuits Syst I* 47(7):1109–1114
- Suykens JAK, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* 48(1–4):85–105. [https://doi.org/10.1016/S0925-2312\(01\)00644-0](https://doi.org/10.1016/S0925-2312(01)00644-0)
- Taormina R, Chau K, Sethi R (2012) Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Eng Appl Artif Intell* 25(8):1670–1676. <https://doi.org/10.1016/j.engappai.2012.02.009>
- Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
- Wang B, Oldham C, Hipsey MR (2016) Comparison of machine learning techniques and variables for groundwater dissolved organic nitrogen prediction in an urban area. *Proc Eng* 154:1176–1184. <https://doi.org/10.1016/j.proeng.2016.07.527>
- Wunsch A, Liesch T, Broda S (2018) Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). *J Hydrol* 2(2):1–15. <https://doi.org/10.1016/j.jhydrol.2018.01.045>
- Yoon H, Jun S-C, Hyun Y, Bae G-O, Lee K-K (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J Hydrol* 396(1–2):128–138. <https://doi.org/10.1016/j.jhydrol.2010.11.002>
- Yoon H, Hyun Y, Ha K, Lee K-K, Kim G-B (2016) A method to improve the stability and accuracy of ANN- and SVM-based time series models for long-term groundwater level predictions. *Comput Geosci* 90:144–155. <https://doi.org/10.1016/j.cageo.2016.03.002>
- Zhang L, Suganthan PN (2015) Oblique decision tree ensemble via multi-surface proximal support vector machine. *IEEE Trans Cybern* 45(10):2165–2176

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