

RESEARCH

Open Access



# An advanced hybrid deep learning model for predicting total dissolved solids and electrical conductivity (EC) in coastal aquifers

Zahra Jamshidzadeh<sup>1</sup>, Sarmad Dashti Latif<sup>2,3</sup>, Mohammad Ehteram<sup>4</sup>, Zohreh Sheikh Khozani<sup>5</sup>, Ali Najah Ahmed<sup>6\*</sup>, Mohsen Sherif<sup>7</sup> and Ahmed El-Shafie<sup>8</sup>

## Abstract

For more than one billion people living in coastal regions, coastal aquifers provide a water resource. In coastal regions, monitoring water quality is an important issue for policymakers. Many studies mentioned that most of the conventional models were not accurate for predicting total dissolved solids (TDS) and electrical conductivity (EC) in coastal aquifers. Therefore, it is crucial to develop an accurate model for forecasting TDS and EC as two main parameters for water quality. Hence, in this study, a new hybrid deep learning model is presented based on Convolutional Neural Networks (CNNE), Long Short-Term Memory Neural Networks (LOST), and Gaussian Process Regression (GPREG) models. The objective of this study will contribute to the sustainable development goal (SDG) 6 of the United Nations program which aims to guarantee universal access to clean water and proper sanitation. The new model can obtain point and interval predictions simultaneously. Additionally, features of data points can be extracted automatically. In the first step, the CNNE model automatically extracted features. Afterward, the outputs of CNNE were flattened. The LOST used flattened arrays for the point prediction. Finally, the outputs of the GPREG model receive the outputs of the LOST model to obtain the interval prediction. The model parameters were adjusted using the rat swarm optimization algorithm (ROSA). This study used PH, Ca<sup>++</sup>, Mg<sup>2+</sup>, Na<sup>+</sup>, K<sup>+</sup>, HCO<sub>3</sub>, SO<sub>4</sub>, and Cl<sup>-</sup> to predict EC and TDS in a coastal aquifer. For predicting EC, the CNNE-LOST-GPREG, LOST-GPREG, CNNE-GPREG, CNNE-LOST, LOST, and CNNE models achieved NSE values of 0.96, 0.95, 0.92, 0.91, 0.90, and 0.87, respectively. Sodium adsorption ratio, EC, magnesium hazard ratio, sodium percentage, and total hardness indices were used to evaluate the quality of GWL. These indices indicated poor groundwater quality in the aquifer. This study shows that the CNNE-LOST-GPREG is a reliable model for predicting complex phenomena. Therefore, the current developed hybrid model could be used by private and public water sectors for predicting TDS and EC for enhancing water quality in coastal aquifers.

**Keywords** Water quality, Sustainable development goal (SDG), Deep learning, Optimization algorithms, Point prediction, Interval prediction

\*Correspondence:

Ali Najah Ahmed  
ale.najah@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Introduction

Coastal freshwater aquifers offer water for a variety of vital uses, including municipal and domestic water supplies, crop and pasture irrigation, and industrial activities. The coastal aquifer (CA) is an important natural resource for socioeconomic development [15]. The water quality of coastal aquifers depends on several factors, including climate change, population growth, geological formations, and recharge rates. The water quality directly affects public health and the environment [3]. Monitoring and evaluating the water quality of coastal aquifers is essential because they are used for irrigation and drinking [35]. Predicting the water quality of coastal aquifers helps decision-makers to reduce pollution. Conventional methods of assessing water quality are usually expensive and time-consuming for decision-makers, especially in developing countries [10]. Water quality can be predicted and managed using various physical or mathematical models. However, these models are complex, time-consuming, and data-intensive [29]. It is difficult to use these models in developing countries due to the insufficiency of data or a scarcity of background information.

Various soft computing models have been used to predict water quality over the past few years [28, 22, 21, 43]. In order to predict water quality parameters, machine learning models are a better choice than sensors because of the following reasons:

1. **Accuracy:** Machine learning models can provide more accurate predictions than sensors [5]. Machine learning models can analyze complex data patterns and make predictions based on them.
2. **Scalability:** Machine learning models can be trained on large volumes of data, so they can predict water quality parameters across different regions and time periods. Sensors have a limited range of applications and may not be able to collect data from multiple locations [8].
3. **Flexibility:** Machine learning models can adapt to different water quality parameters, making them more versatile than sensors designed for particular parameters. In other words, machine learning models can be customized to meet a variety of needs related to water quality monitoring.
4. **Cost-effective:** Machine learning models are more cost-effective than sensors. Sensors are expensive to deploy and maintain.
5. **Reliability:** Machine learning models are more reliable than sensors, which may malfunction or be affected by environmental factors [5]. When sensors fail or are unavailable, machine learning models can still provide accurate predictions.

Various research has been conducted to determine and forecast groundwater level [26, 27]. For instance, for predicting the electrical conductivity (EC) of groundwater, Khashei-Siuki et al. [18] used the kriging method, artificial neural networks (ANNs), and adaptive neuro-fuzzy inference systems (ANFISs). A high correlation was found between the  $\text{Cl}^-$  and EC parameters. ANN showed the best accuracy. Ravansalar and Rajaei [31] developed an ANN and wavelet ANN model to predict the monthly EC. Their results indicated that wavelet ANN was superior to ANN. Mohammadpour et al. [25] used radial basis function neural networks (RBFNNs) and support vector machine (SVM) models to predict the water quality index. Based on their study, SVMs and RBFNNs could successfully predict water quality indexes. Using wavelet-ANFIS and wavelet-ANN, Barzegar et al. [7] predicted electrical conductivity-based salinity levels.  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{SO}_4^{2-}$ , and  $\text{Cl}^-$  were the inputs. Wavelet-ANFIS outperformed the Wavelet-ANN model. Salami et al. [33] used ANN models to predict dissolved oxygen (DO) and total dissolved solids (TDS). The ANN models were reliable for predicting water quality indicators. Amanollahi et al. [2] evaluated the ability of remote sensing data to predict TDS and PH using. The ANN model and remote sensing data successfully predicted water quality indicators. Charulatha et al. [9] used principal component regression (PCR)-ANN to estimate nitrite concentration. For predicting nitrite concentrations, the PCR-ANN showed high potential. For predicting DO, Zhang et al. [40] used an SVM model. The authors proposed a particle swarm optimization algorithm (PSOA) for finding SVM parameters. They concluded that SVM-PSO was a robust tool for short-term prediction. Khadr and Elshemy [17] used the ANFIS model to predict total phosphorus and nitrogen. ANFIS model required inputs such as TDS, EC, and PH. As a predictive tool, they found the ANFIS model to be reliable. Ahmed and Shah [1] used the ANFIS model to estimate DO. The ANFIS model was reliable for predicting water quality indicators. For EC prediction, Barzegar et al. [8] used extreme Learning Machine (ELM) models and wavelet-ELMs. The least squares boosting (LSBoost) algorithm was used to create an ensemble model based on the outputs of ELM and wavelet-ELM models. The ensemble model outperformed the wavelet-ELM and ELM models. Zhu and Heddami [42] used ANN and ELM models to predict DO. Overall, the ELM and ANN models successfully predicted DO. For predicting the water quality index, Kouadri et al. [19] suggested ANN, multilinear regression (MLR), and support vector machines (SVM). These models had high abilities for predicting the water quality index in the study area. Azrour et al. [4] used ANN

and multiple regression algorithms to predict the water quality index. They stated that the ANN and MLR successfully predicted the water quality index. SVM, ELM, MLP, RBFNN, and ANFIS have successfully been used for predicting water quality. However, these models have some shortcomings. These models may miss information in the modeling process. These models can not automatically extract the features of input data.

Deep learning (DL) models are widely used to address the shortcomings of soft computing models. Deep learning models can extract deep features from data points. A convolutional neural network (CNN) is one of the robust deep learning models. CNN has been widely used in different fields, such as medical image [34], prediction of plant leaf diseases [12], stock trend prediction [11], streamflow prediction [14], and weather radar echo prediction [14]. A CNN model can extract data features, but it may not be able to learn sequence associations. Due to their excellent information memory and sequential modeling capabilities, long short-term memory (LSTM) networks are used for simulating complex problems [30, 38]. Hence, CNNE-LSTM models are suggested for extracting complex features and predicting outputs. A CNNE-LSTM combines the advantages of CNNEs and LSTMs. For time series data, the LSTM has excellent processing ability, while the CNNE extracts features of grid data. Kumari and Toshniwal [20] used LSTM-CNNE models to predict global horizontal irradiance. They reported that the LSTM-CNNE model was a robust tool for short-term predictions. Yan et al. [39] used CNNE-LSTM models to predict air quality. They reported that the LSTM-CNNE outperformed the LSTM and CNN models.

However, CNNE-LSTM only provides a single prediction value. During the modeling process, it is essential to obtain the interval prediction and uncertainty values. Systematic reviews have shown that Gaussian process regression (GPR) is a useful method for interval prediction [36, 37]. GPR is a type of nonlinear Bayesian regression for quantifying uncertainty.

Using LSTM and CNN, features can be extracted from the input data. Then, the GPR is used to provide reliable interval predictions. A CNNE-LSTM-GPR can predict points as well as intervals simultaneously. There are various advantages of the current developed hybrid model. For instance, the CNNE-LSTM-GPR model predicts both interval and point predictions simultaneously. Secondly, unlike MLP, RBFNN, and SVM models, the CNNE-LSTM-GPR extracts features automatically. Finally, it is possible to quantify the uncertainty of the modeling process using CNNE-LSTM-GPR.

Hence, this study introduces the new hybrid model, namely, CNNE-LSTM-GPR for predicting TDS and EC in a coastal aquifer. EC and TDS are predicted because

they are the most important water quality indicators. Predicting the electrical conductivity of water provides valuable information about its purity or contamination. The electrical conductivity of water is directly related to the dissolved ions or salts in the water. Higher electrical conductivity in water indicates more dissolved solids, which can negatively impact aquatic life, human health, and industrial processes. A lower electrical conductivity indicates lower levels of contamination and higher purity of water, making it safe for consumption. Therefore, predicting the electrical conductivity of water is important to monitor and regulate water quality and ensure ecosystem health.

## Material and method

### Structure of convolutional neural network models (CNN)

Because CNNE models share feature parameters and reduce dimensionality, they are widely used for predicting outputs [36]. By sharing parameters, CNNE reduces the number of parameters and computations. CNNE consists of convolutional, pooling, and fully connected layers [6]. The convolutional layer consists of many convolution kernels. From input matrices, convolution kernels generate feature maps. Spatial and temporal dependencies are captured using the convolution kernels. A pooling layer decreases the spatial dimensions of the matrices by down-sampling them. In the pooling layer, the number of parameters is reduced while the essential characteristics are maintained. Through fully connected layers, latent patterns are learned from time series input, feature maps, and targets. CNNEs commonly use Rectifying Linear Activation Units (ReLU) as activation functions. In this study, the weight connections of the CNNE are updated using a robust optimization algorithm.

### Structure of LSTM

LSTM is a robust method for sequence learning. A LSTM has a memory cell that can retain information for a long period. There are three multiplicative units in each layer: input gate, forget gate, and output gate. LSTM uses state cells. Using the forget gate, it is possible to determine what information should be removed or wished for [41].

$$f_t = \mu(\omega_f \cdot [h_{t-1}, x_t] + \beta_f) \quad (1)$$

where  $f_t$ : the activation values of the forget gate  $\omega_f$ : the weight matrix of the forget gate,  $\beta_f$ : the bias matrix of the forget gate, and  $\mu$ : the activation function. Input gates determine what information is added to a cell state. The process consists of two levels. The first step is calculating candidate values for the cell states [23]. The next step is to calculate the activation values of the input gates.

$$\tilde{\rho}_t = \tanh(\omega_\rho \cdot [h_{t-1}, x_t] + \beta_p) \quad (2)$$

$$i_t = \mu(\omega_i \cdot [h_{t-1}, x_t] + \beta_i) \quad (3)$$

where  $\omega_\rho$  and  $\omega_i$ : the weight matrices of cell state and input gate,  $\beta_i$  and  $\beta_p$ : bias matrix,  $\tilde{\rho}_t$ : candidate values for the cell states,  $x_t$ : input,  $h_{t-1}$ : hidden state, and  $i_t$ : activation values of the input gates. Based on the previous levels, new cell states are computed.

$$\rho_t = f_t * \rho_{t-1} + \tilde{\rho}_t \quad (4)$$

where  $\rho_t$ : cell state at time t, and  $\rho_{t-1}$ : cell state at time t-1. Finally, the output gate provides the outputs:

$$o_t = \mu(\omega_o \cdot [h_{t-1}, x_t] + \beta_o) \quad (5)$$

$$h_t = o_t \tanh(\rho_t) \quad (6)$$

where  $o_t$ : activation values of the input gates,  $\omega_o$  and  $\beta_o$ : weight and bias matrices of output gate  $h_t$ : output.

### Structure of Gaussian process regression (GPRE)

GPR is a nonparametric probabilistic model for quantifying uncertainty [16]. GPRE is a good choice for approximating nonlinear functions. For the noisy data, a regression model is considered as follows:

$$Z = f(in) + v \quad (7)$$

where  $Z$ : output,  $f$ : basic function  $in$ : input, and  $v$ : noise. Then, the prior distribution of observed data can be computed.

$$Z \sim N\left(0, K(in_i, in_j) + \sigma_n^2 I_n\right) \quad (8)$$

where  $\sigma_n^2$ : variance,  $I_n$ : unit matrix,  $in_i$ :  $i$ th input,  $in_j$ :  $j$ th input, and  $K(in_i, in_j)$ : the  $N$ -dimensional covariance matrix. The covariance matrix is computed as follows [37]:

$$K(in_i, in_j) = \text{cov}(in_i, in_j) = \sigma_f^2 \exp\left(-\frac{(in_i - in_j)^2}{l^2}\right) \quad (9)$$

where  $\sigma_f$  and  $l$ : hyperparameters. Lastly, the posterior distribution of the predicted value is calculated.

$$z|Z \sim N(\bar{z}, \sigma_z^2) \quad (10)$$

$$\bar{z} = K_* K^{-1} Z \quad (11)$$

$$\sigma_z^2 = K_{**} - K_* K^{-1} K_*^T \quad (12)$$

where  $K_{**}$ : the self-covariance of test points,  $K_*$ : the  $n^* \times 1$  covariance matrix of test points,  $z$ : the point prediction results of GPR, and  $\sigma_z^2$ : variance of the predicted value. Since the CNN-LOST model gives the point predictions, we only require  $\sigma_z^2$  to obtain the corresponding interval prediction (CIP) ( $\bar{z} - 1.96 \sigma_z$ ,  $\bar{z} + 1.96 \sigma_z$ ). The following equation computes the probability density function of the predicted value:

$$p(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(Z_i - \bar{z}_i)^2}{2\sigma_{z_i}^2}\right) \quad (13)$$

### The structure of RSOA

There are many optimization algorithms, but RSOA is a simple and robust algorithm for solving complex problems. Based on the life of rats, Dhiman et al. [13] introduced RSOA. Rats are aggressive animals that can kill their enemies through their aggressive behavior. For solving complex problems, the RSO mathematically simulates the chasing and fighting behaviors of rats. Generally, chasing behavior assumes that the best search agent knows the location of prey before beginning its search. Based on the location of the best search agent, the other rats update their locations. Using the following equation, we can simulate chasing behavior [13]

$$\vec{R}\vec{A} = A \cdot \vec{R}_i(x) + C \cdot (\vec{R}_r(x) - \vec{R}_i(x)) \quad (14)$$

$$A = \alpha - IT \left( \frac{\alpha}{IT_{\max}} \right) \quad (15)$$

$$C = 2 \cdot \text{rand} \quad (16)$$

where  $\vec{R}_i(x)$ : the current location of rats,  $\vec{R}_r(x)$ : The best location of rats,  $A$  and  $C$ : random parameters,  $\text{rand}$ : random number,  $IT$ : number of iterations,  $IT_{\max}$ : maximum number of iterations,  $\alpha$ : constant value,  $\vec{R}\vec{A}$ : the updated location of rats and  $C$ : random numbers. At the net level, the following equation is used to simulate the fighting behavior of RSOA:

$$\vec{R}\vec{A}_i(x+1) = \left| \vec{R}\vec{A}_r(x) - \vec{R}\vec{A}_i \right| \quad (17)$$

where  $\vec{R}\vec{A}_i(x+1)$ : the new position of the rat.

### Structure of hybrid LOST-RSO, CNNE-RSO, and CNNE-LOST-GPRE

Weight and bias are the key parameters of LOST and CNNE models. In this study, the RSO was used to adjust the LOST and CNNE parameters:

- 1) For LOSTEs and CNNEs, weights and biases are initialized.
- 2) A CNNE and a LOST are run using training data.
- 3) Check the stop criterion (CC). Models are run at the testing level if CC is met; otherwise, they go to step 4.
- 4) The LOST and CNNE parameters are regarded as the initial population of the algorithms.
- 5) Each rat's location represents the weight and bias parameter values.
- 6) The models are run using the initial population of the algorithms.
- 7) The objective function (root mean square error) assesses the quality of the solution.
- 8) Equations 16 and 17 are used to update rat locations using the operators of rat algorithms.
- 9) The models go to step 3 if the convergence criterion is met; otherwise, they go to step 6.

CNNE-LOST-GPR is a hybrid model for predicting complex phenomena. Each model has a task in the modeling process. Training data are inserted into the CNNE model in the first step. The convolutional layer (COL) extracts features using convolution kernels. COLs provide feature maps. A pooling layer decreases the width and length of feature maps. Finally, CNNE provides outputs. In the next level, these outputs are flattened. The flattened arrays are inserted into the LOST model. Figure 1 demonstrates the structure of the LOST-CNNE model. The LOST model provided point predictions at the training and testing levels. Then, the outputs of LOST models are inserted into the GPR model for interval predictions. The GPRE predicts all data points and obtains interval predictions. This study compares CNNE-LOST-GPRE with LOST-CNNE, LOST, CNNE, LOST-GPRE, and CNNE-GPRE models. The structure of hybrid models is explained based on the following levels.

- Hybrid CNN-LOST

CNNE extracts the feature at the training and testing levels. The flattened outputs of CNNE are inserted into the LOST model for predicting data points.

- Hybrid CNNE-GPRE

The training and test data were inserted into the CNNE model at the training and testing levels. The outputs of the CNNE model are flattened. The flattened outputs are inserted into the GPRE model. The GPRE model provides interval predictions.

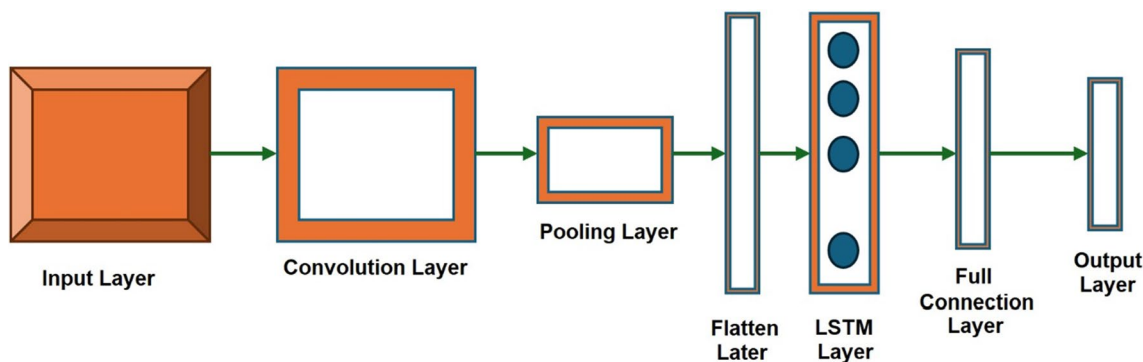
- Hybrid LOST-GPR

The training and testing data were used to run the LOST model at the training and testing level. The outputs of the LOST model are inserted into the GPRE model for interval predictions.

For predicting TDS, the daily inputs were PH,  $\text{Ca}^{++}$ ,  $\text{Mg}^{++}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{HCO}_3$ ,  $\text{SO}_4$ , and  $\text{Cl}^-$  and for predicting EC, the inputs were PH,  $\text{Ca}^{++}$ ,  $\text{Mg}^{++}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{HCO}_3$ ,  $\text{SO}_4$ , and  $\text{Cl}^-$ .

### Case study

This paper studies Ghaemshahr coastal aquifer which is located in the north of Iran. A dense forest surrounds the southern region of the basin, while the Caspian Sea surrounds the northern part. There are sub-humid and humid climates in the region. In the study area, 85% of groundwater is used for agricultural purposes. Additionally, groundwater meets about 75% of drinking water demands. Therefore, the plain plays a key role in the water supply. River deposits have formed several types of alluvial plains within the study area. The



**Fig. 1** Structure of the LOST-CNNE model

shallow unconfined aquifer was formed by a calcareous unit containing sand and gravel. Silty and clayey sediments separate the semi-confined aquifer from the unconfined aquifer. The percolated rainfall dissolves minerals in the recharge zone due to the presence of calcareous and dolomite rocks. The data were collated from three zones and observed well.

In zone A (the recharge zone near the foothills of the alborz mountains), the groundwater table level changes from 55 (at sampling point 15) to 94 m (at sampling point 2) above the Caspian Sea level. Water well depth within zone A ranges from 21 to 187 m below the ground surface. In this zone, both the underlying semi-confined and the top unconfined aquifers are connected hydraulically and operate as a unified aquifer system. Water table level in zone B (the central zone) composed of stratified sediments (the top unconfined aquifer), the aquitard layer, the semi-confined aquifer, and the marine sediments) range between 6.6 (sampling point 29) and 61.7 m (sampling point 33) above the Caspian Sea level. Zone C is located near the coastline, and the water table level ranges from 0.4 (sampling point 53) to 12.4 m (sampling point 68) above the mean level of the Caspian Sea. Water wells in this zone are at shallow depths ranging from 12 to 24 m from the ground level.

The study period is from 2015 to 2021. For predicting TDS, the daily inputs were PH, Ca<sup>++</sup>, Mg<sup>++</sup>, Na<sup>+</sup>, K<sup>+</sup>, HCO<sub>3</sub>, SO<sub>4</sub>, and Cl<sup>-</sup> and for predicting EC, the inputs were PH, Ca<sup>++</sup>, Mg<sup>++</sup>, Na<sup>+</sup>, K<sup>+</sup>, HCO<sub>3</sub>, SO<sub>4</sub>, and Cl<sup>-</sup>. Table 1 shows the statistical details of input and output data. Figure 2 shows the study area on Google Map while Fig. 3 shows data points of EC and TDS while.

In some points of Fig. 3, the EC is very high due to various factors. For instance, when the temperature decreases, the EC will increase due to decreasing electrons scattering. Moreover, type and concentrations of

ions are also another factor that affects the changes in EC.

In this study, point prediction evaluation metrics are applied to evaluate the performance of models:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (V_i - v_i)^2} \tag{18}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |(V_i - v_i)| \tag{19}$$

$$NSE = 1 - \frac{\sum_{i=1}^N (v_i - V_i)^2}{\sum_{i=1}^N (V_i - \bar{V})^2} \tag{20}$$

$$PBIAS = \frac{\sum_{i=1}^N (V_i - v_i)}{\sum_{i=1}^N (V_i)} \tag{21}$$

where MAE mean absolute error, RMSE: root mean square error, N: number of data, V<sub>i</sub>: Observed data,  $\bar{V}_i$ : average observed data, v<sub>i</sub>: estimated data,  $\bar{v}_{ies}$ : average estimated data, PBIAS: Percent bias, and NSE: Nash–Sutcliffe efficiency. The low values of RMSE, MAE, and PBIAS show the best efficiency. The following indices are used to evaluate the predicted intervals:

$$PICP = \frac{1}{N} \sum_{i=1}^N \rho_i \tag{22}$$

$$\rho_i = \begin{cases} 1 & \leftarrow \text{if } (low_i) \leq v_i \leq (up_i) \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

$$PINAW = \frac{1}{NR} \sum_{i=1}^N (up_i - low_i) \tag{24}$$

$$NC = \frac{PINW}{PICP} \tag{25}$$

where PICP: Prediction Interval Coverage Probability, N: number of data, R: range of data, PINW: Prediction Interval Normalized Average Width, up<sub>i</sub>: upper values of variables, and low<sub>i</sub>: lower values of variables, NC: index uncertainty. The low and high values of PINAW and PICP show more accurate predictions. Table 2a, b show the optimal values of model parameters.

## Results and discussions

### Selection of the size of data

The optimal size of the training and testing sets are selected based on the individual models. For instance, for the hybrid CNN-LOST model, CNNE extracts the feature

**Table 1** The details of input and output data (number of input data:391, number of output data:391)

| Parameter        | Maximum | Average | Minimum |
|------------------|---------|---------|---------|
| PH               | 8.4     | 7.1     | 6.7     |
| TDS (mg/lit)     | 2818.2  | 1354.4  | 312.0   |
| EC (μS/cm)       | 4310.0  | 2200.6  | 551.00  |
| Mg <sup>++</sup> | 86.2    | 22.12   | 5.1     |
| Ca <sup>++</sup> | 211.2   | 112.2   | 27.2    |
| SO <sub>4</sub>  | 437.2   | 55.23   | 5.21    |
| HCO <sub>3</sub> | 871.2   | 404.2   | 55.12   |
| K <sup>+</sup>   | 5.2     | 3.32    | 1.23    |
| Na <sup>+</sup>  | 723.2   | 131.2   | 9.4     |



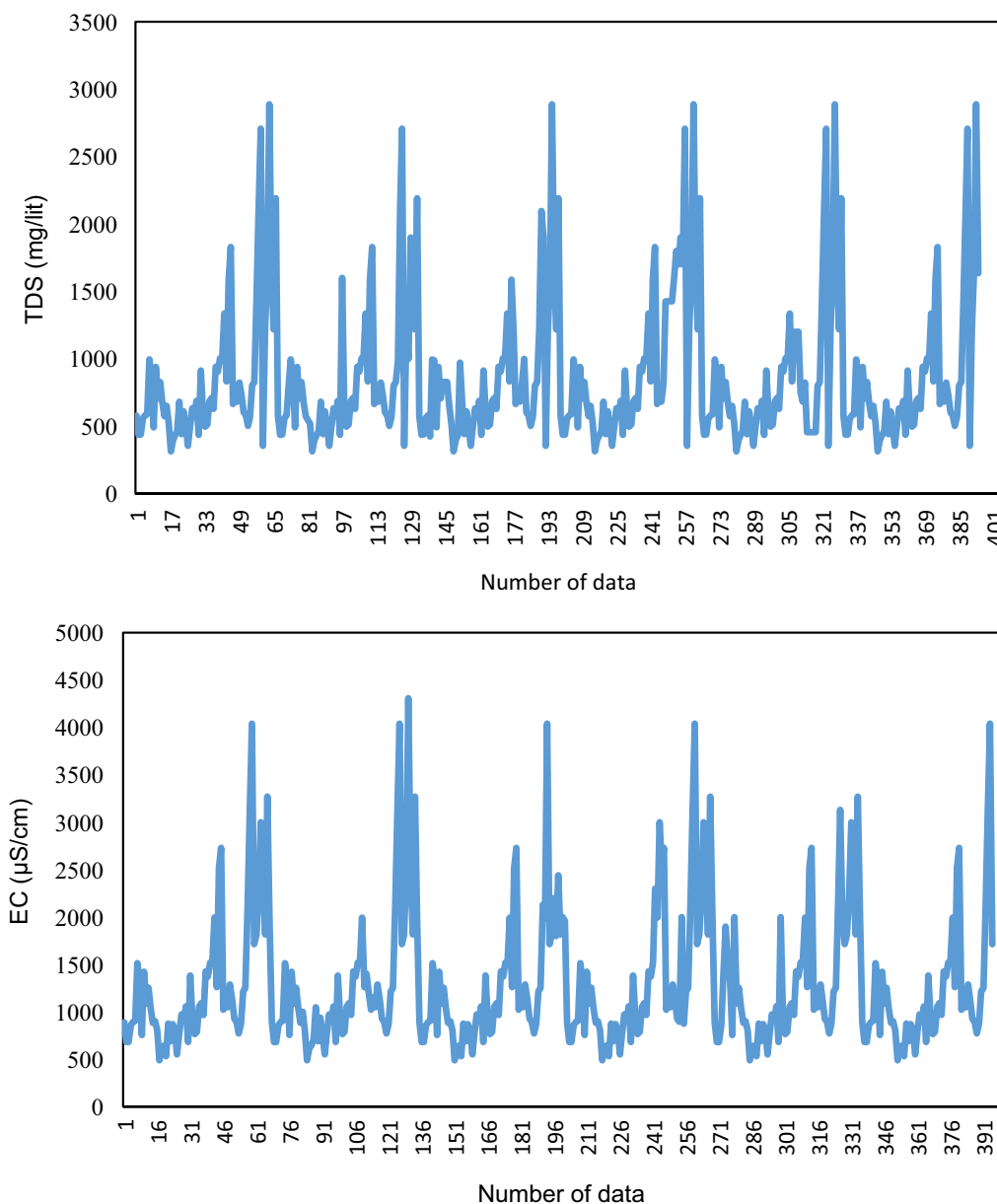
**Fig. 2** Study area on Google Map

at the training and testing levels. The flattened outputs of CNNE are inserted into the LOST model for predicting data points. Therefore, each model uses different sizes for training and testing sets. Based on different data sizes, Fig. 4 shows the RMSE values of CNNE-LOST-GPRE. For predicting EC, the RMSEs of 50, 55, 60, 65, 70, 75, 80, and 85% of data were 10.00, 7.0, 2.2, 5.1 mg/lit, 6.0, 7.0, 8.0, and 8.3 mg/lit. For predicting TC, the RMSEs of 50, 55, 60%, 65%, 70%, 75%, 80%, and 85% of data were 9.00 mg/lit, 8.0 mg/lit, 2.5 mg/kit, 5.4 mg/lit, 6.2 mg/lit, 7.1 mg/lit, 8.0 mg/lit, and 8.7 mg/lit.

#### Determination of random parameters

The performance of RSOA depended on the values of random parameters. Therefore, it is necessary to

determine the values of random parameters. The maximum number of iterations (MANU) and population size (POPS) are the two most important parameters of RSOA. MANU and POPS are calculated using sensitivity analysis in this study. Minimizing the objective function is obtained by adjusting parameter values. Therefore, the lowest values of random parameters gave the lowest values of the objective function. Figure 5 shows a heat map for determining parameters. For EC prediction, the RMSEs of MANU=150, MANU=300, MANU=450, MANU=600, and MANU=750 were 9.4 mg/lit, 2.5 mg/lit, 6.8 mg/lit, 7.9 mg/lit, and 8.3 mg/lit, respectively. For TDS prediction, objective function (RMSE) values of the MANU=150, MANU=300, MANU=450, MANU=600, and MANU=750 were 9.5 mg/lit, 2.4 mg/



**Fig. 3** Data points of EC and TDS

lit, 3.2 mg/lit, 4.5 mg/lit, and 5.8 m/lit, respectively. Thus, MAENU=300 provided the lowest value of the objective function (OBF). For EC predictions, the objective function (OBF) values of POPS=65, POPS=130, POPS=195, POPS=260, and PSOP=325 were 9.2, 2.3, 4.8, 6.8, and 8.2, respectively. For TDS prediction, the OBF values of POPS=65, POPS=130, POPS=195, POPS=260, and PSOP=325 were 9.3, 2.5, 3.1, 4.7, and 5.9, respectively.

**Selected features by the hybrid model**

This study uses hybrid GPR-CNN-LOST to identify features automatically. The best input combinations are shown in Table 3. For predicting TDS, the best input combination was HCO<sub>3</sub>, Na<sup>+</sup>, Ca<sup>++</sup>, and Mg<sup>++</sup>. For Predicting EC, the best input combination was Na<sup>+</sup>, HCO<sub>3</sub>, SO<sub>4</sub>, and Ca<sup>++</sup>. However, it is necessary to evaluate the performance of hybrid GPRE-CNNE-LOST models



**Table 2** Optimal values of model parameters, a: for predicting EC, and b: for predicting TDS

| a               |   |
|-----------------|---|
| Model           | Type of activation function hyperparameter  |
| LOST            | Number of hidden layers:8, fixed-rate learning: 0.01, size of the batch: 32, and epochs of training: 1000   |
| GPRE            | Kernel function: Gaussian function, $\sigma_f^2$ :2 and l:1   |
| CNNE            | learning rate:0.01, kernel size:2 and pooling size:1  |
| LOST-GPRE-CNNE: | LOST (Number of hidden layers:5, fixed-rate learning: 0.01, size of the batch: 32, and epochs of training: 1000)<br>CNN (Learning rate:0.01, kernel size:2, and pooling size:1)<br>GPR (kernel function: Gaussian, $\sigma_f$ :2 and l:1) |
| b               |   |
| Model           | Parameter values  |
| LOST            | Number of hidden layers:8, fixed-rate learning: 0.01, size of the batch: 32, and epochs of training: 2000   |
| GPRE            | Kernel function: Gaussian function, $\sigma_f^2$ :2 and l:1   |
| CNNE            | learning rate:0.01, kernel size:2 and pooling size:1  |
| LOST-GPRE-CNNE: | LOST (Number of hidden layers:8, fixed-rate learning: 0.01, size of the batch: 32, and epochs of training: 1000)<br>CNN (Learning rate:0.01, kernel size:2, and pooling size:1)<br>GPR (kernel function: Gaussian, $\sigma_f$ :2 and l:1) |

when selecting features. Previous research showed the effect of  $\text{HCO}_3$  on EC [32]. Figure 6 indicates the correlation heat maps between outputs and inputs. It was found that  $\text{HCO}_3$ ,  $\text{Na}^+$ ,  $\text{Ca}^{++}$ , and  $\text{Mg}^{++}$  had the highest correlation with TDS. It was found that  $\text{Na}^+$ ,  $\text{HCO}_3$ ,  $\text{SO}_4$ , and  $\text{Ca}^{++}$  had the highest correlation with EC. Thus, the hybrid model correctly chooses the best features. Also, LOST, GPRE, CNNE-, LOST-CNNE, LOST-GPRE, and CNNE-GPRE used the best input combinations for predicting TDS and EC.

The correlation heat maps between outputs and inputs have been clearly shown in Fig. 6. For instance, the correlation values for pH are 0.3 and 0.59 for input and output of TDS respectively. Moreover, the correlation values for pH are 0.54 and 0.73 for input and output of EC respectively.

#### Evaluation of the accuracy of models for point predictions

This section evaluates the accuracy of models for predicting points.

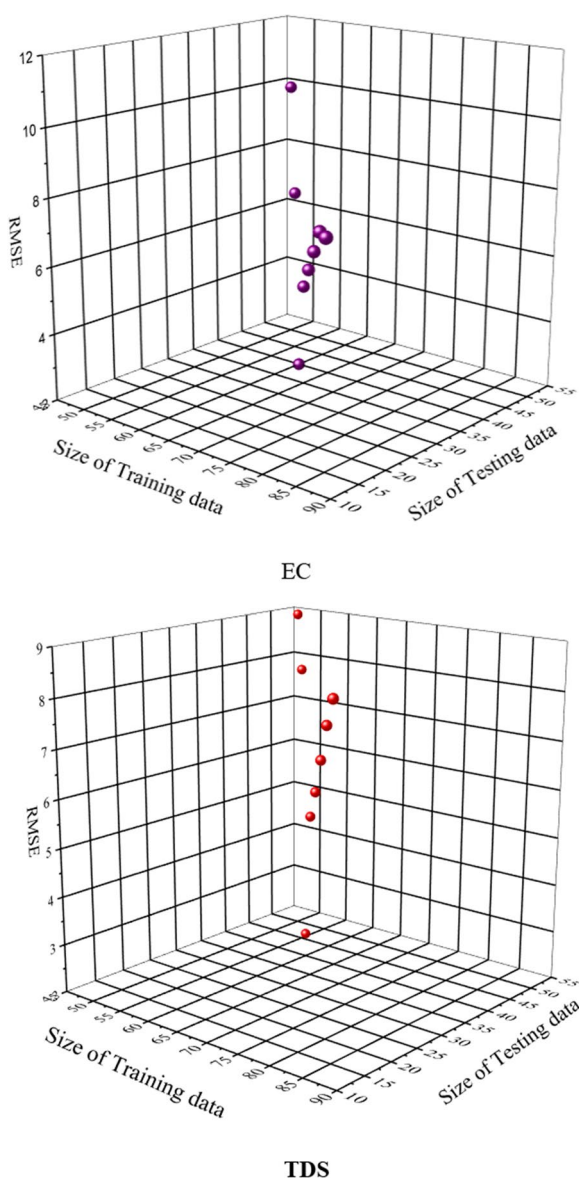
- EC

Figure 7 shows values of error indices for EC prediction. At the training level, the MAEs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST,

and CNNE model were 1.67, 1.75, 1.9, 2.35, 3.24, and 4.25 mg/lit, respectively (Fig. 7). The CNN-LOST-GPR decreased the MAE of the LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models by 12, 14, 27, 50, and 64%, respectively. The training NSEs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 0.98, 0.97, 0.94, 0.93, 0.92, and 0.89, respectively. The testing NSEs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 0.96, 0.95, 0.92, 0.91, 0.90, and 0.87, respectively. The training PBIASs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, and CNNE models were 4, 7, 9, 11, 12, and 14, respectively. At the testing level, the PBIASs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST and CNNE models were 5, 8, 11, 12, 14, and 15, respectively. The radar plots of error indices are shown in Figs. 6, 7.

- TDS

Figure 8 shows values of error indices for EC prediction. The training MAEs of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-LOST, LOST, and CNNE model were 1.55, 1.73, 1.88, 2.21, 3.29, and 4.22 mg/lit, respectively. The CNN-LOST-GPR decreased the testing



**Fig. 4** The RMSE values for different data sizes

MAEs of the LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models by 2.1, 12, 24, 48, and 60%, respectively. The training NSE values of the CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 0.97, 0.95, 0.93, 0.92, 0.90, and 0.88, respectively. The testing NSEs of the CNNE-LOST-GPRE, LOSTE-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 0.95, 0.94, 0.91, 0.90, 0.89, and 0.87, respectively. The training PBIAS values of the CNNE-LOST-GPRE, LOSTE-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 3, 5, 8, 10, 11, and 12, respectively. The testing PBIASs of the CNN-LOST-GPRE, LOST-GPRE, CNNE-GPRE,

CNNE-LOST, LOST, and CNNE models were 6, 7, 9, 11, 13, and 14, respectively.

Figure 9 shows the boxplots of models. A boxplot is a graph that shows how the 25th percentile, 50th percentile, 75th percentile, minimum, maximum, and outlier values of a data set are spread out and compared to one another. The boxplots explain the implemented model for both TDS and EC.

• TDS

The median values of observed data, for models of CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, LOST-CNNE, LOST, and CNNE were 1350, 1350, 1350, 1600, 1650, 1650, and 1750 mg/lit, respectively. The maximum values of observed data for CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, LOST-CNNE, LOST, and CNNE models were 2818, 2898, 2898, 2900, 2923, and 2923 mg/lit. The CNNE-LOST-GPRE and LOST indicated the best and worst performance among other models.

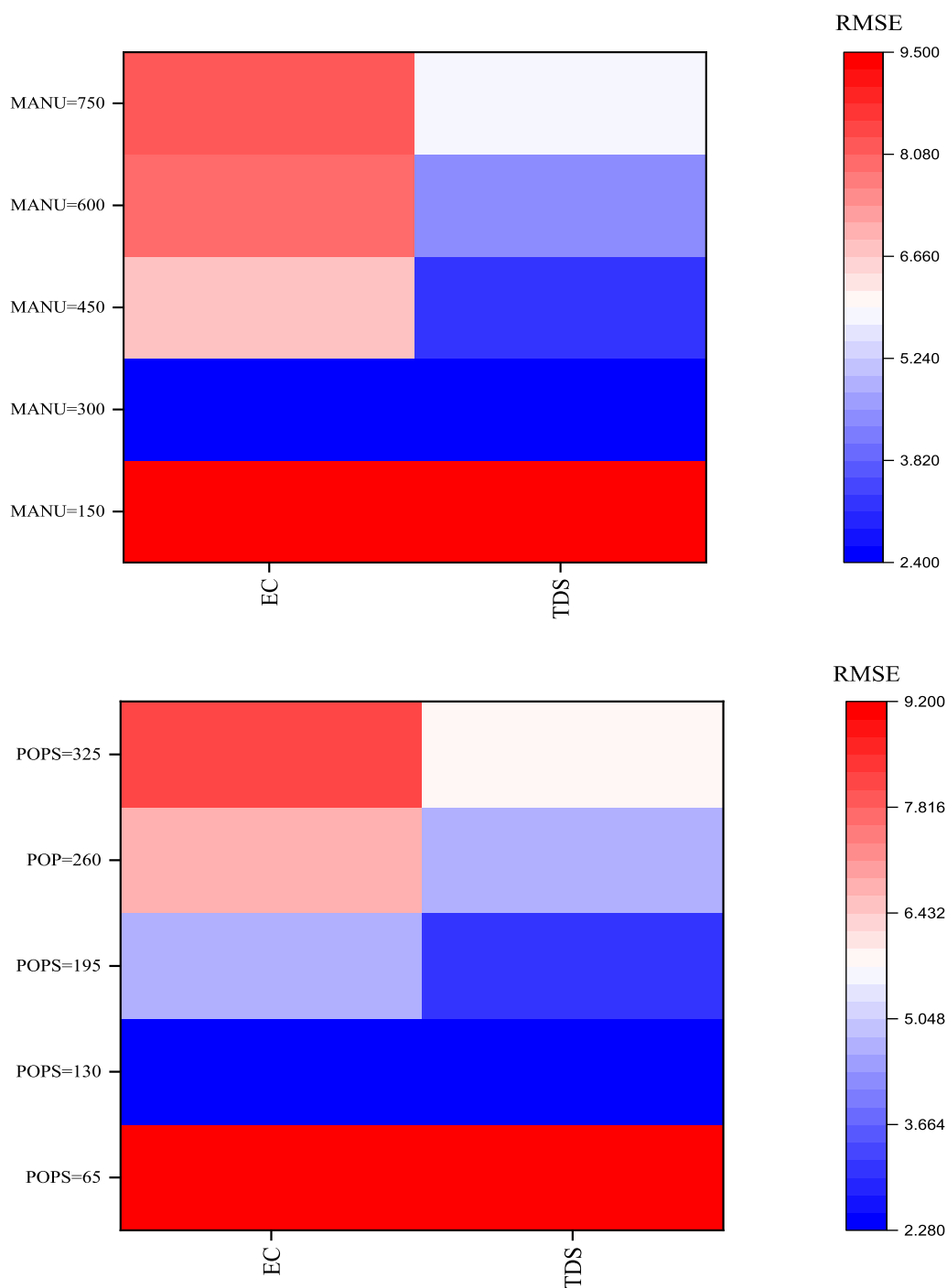
• EC

The median values of observed data, CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, LOST-CNNE, LOST, and CNNE models were 2000 ( $\mu\text{S}/\text{cm}$ ), 2000 ( $\mu\text{S}/\text{cm}$ ), 2000 ( $\mu\text{S}/\text{cm}$ ), 2000 ( $\mu\text{S}/\text{cm}$ ), 2200 ( $\mu\text{S}/\text{cm}$ ), 2300 ( $\mu\text{S}/\text{cm}$ ), and 2400 ( $\mu\text{S}/\text{cm}$ ), respectively. The maximum values of observed data, CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, LOST-CNNE, LOST, and CNNE models were 4310 ( $\mu\text{S}/\text{cm}$ ), 4310 ( $\mu\text{S}/\text{cm}$ ), 4510 ( $\mu\text{S}/\text{cm}$ ), 4545 ( $\mu\text{S}/\text{cm}$ ), 4600 ( $\mu\text{S}/\text{cm}$ ), 4800 ( $\mu\text{S}/\text{cm}$ ), and 4900 ( $\mu\text{S}/\text{cm}$ ). The CNNE-LOST-GPRE and LOST showed the best and worst performance among other models.

**Evaluation of the accuracy of models for interval prediction**

Figure 10 shows the 95% prediction interval for TDS. Prediction interval is the estimation of the interval to fall future observations within certain probabilities. In regression analysis, prediction interval is commonly used. Based on Fig. 10, it can be clearly seen that the extreme events cannot be easily estimated. This is due to the lack of correlation between the previous and next values. The Best performance is achieved when all observed data are within bounds. Models with the highest PICP values are ideal. The CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, and GPRE were used for interval prediction.

The CNNE-LOST GPRE provided the best performance. The PI values of CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, GPRE models were 0.95, 0.94, 0.92,



**Fig. 5** Sensitivity analysis of random parameters of RSOA

**Table 3** The best input combinations for predicting TDS and EC

| Output parameter | The best input combination   |
|------------------|--|
| TDS              | HCO <sub>3</sub> , Na <sup>+</sup> , Ca <sup>++</sup> , and Mg <sup>++</sup> |
| EC               | Na <sup>+</sup> , HCO <sub>3</sub> , SO <sub>4</sub> , and Ca <sup>++</sup>  |

and 0.91, respectively. Figure 11 shows a 95% prediction interval for predicting EC.

The CNNE-LOST-GPRE showed the best performance. The PI values of CNNE-LOST-GPRE, LOST-GPRE, CNNE-GPRE, GPRE models were 0.97, 0.95, 0.93, and 0.90, respectively. Table 4 represents the results of PICP,

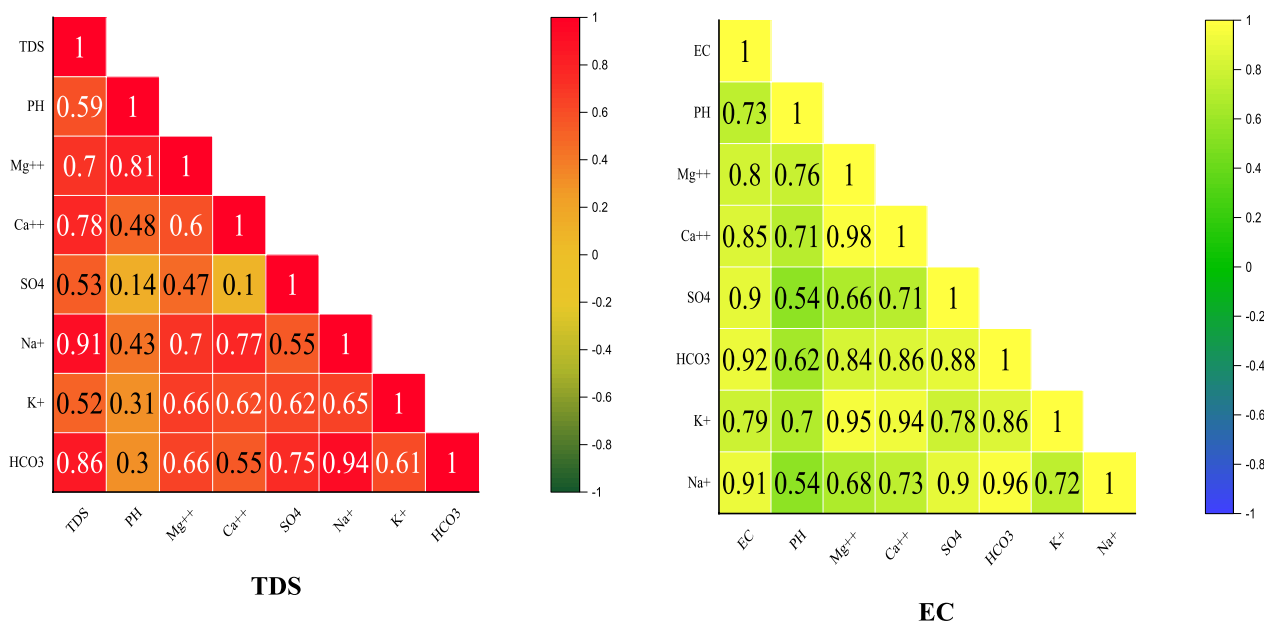


Fig. 6 correlation heat maps

PINW, and NC for both TDS and EC 95% prediction interval.

**Discussion**

**Evaluation of the accuracy of models**

In this study, the CNN-LOST-GPR was used to predict EC and TD. The models were useful for interval and point predictions. The main differences between the current research and other papers were as follows:

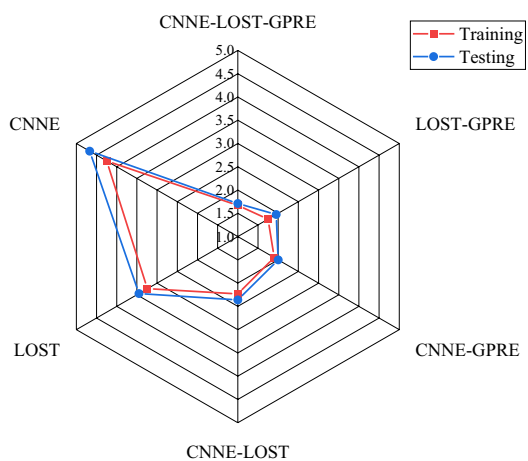
- 1) While previous models, such as MLP, RBFNN, ANFIS, and SVM, could predict points, the new hybrid model could simultaneously predict points and intervals.
- 2) The previous studies used methods such as generalized likelihood estimation for quantifying uncertainty, while the CNNE-LOST-GPRE automatically quantified the uncertainty.
- 3) The previous models, such as MLP, RBFNN, ANFIS, and SVM, needed feature selection methods for choosing features, but the new method automatically selected the features.
- 4) These models can predict other variables such as rainfall, temperature, groundwater level, and streamflow. CNNE models can extract the most important features from different time series. Thus, the modelers can predict outputs best based on input combinations.
- 5) Our study helps improve the accuracy of previous studies. Banadkooki et al. [5] used ANFIS-moth

flame optimization (MFO), ANFIS, and SVM to predict TDS. At the testing, the MAE values of ANFIS-MFO, ANFIS, and SVM were 3.112 mg/lit, 3.186 mg/lit, and 3.238 mg/lit. The MAE of CNNE-LOST-GPRE was 1.79 mg/lit. Thus, CNN-LOST-GPR outperformed the ANFIS-MFO, ANFIS, and SVM models. Mattas et al. [24] used ANN and the multiple linear regression model (MLRM) to predict EC. The NSE values of the MLRM and ANN were 0.94 and 0.93, respectively. The NSE of the CNNE-LOST-GPRE was 0.98 and 0.96 at the training and testing levels. Thus, the CONE-LOST-GPRE outperformed the ANN and MLRM model.

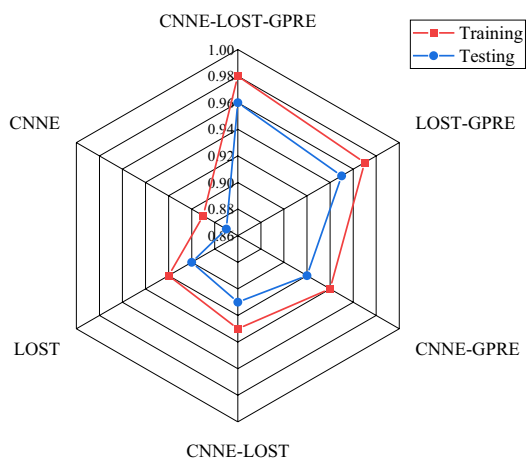
The CNN-LOST-GPR is a robust tool for monitoring water quality in complex and dynamic systems. However, the standalone LOST and CNN were inaccurate in predicting water quality indicators. Also, the high accuracy of CNN-LOST-GPR indicated that the RSOA performed well. The CNN-LOST-GPR also can be used for providing spatial and temporal maps of water quality indicators in a large basin.

**Evaluation of the hadrochemical and water quality characteristics of the aquifer**

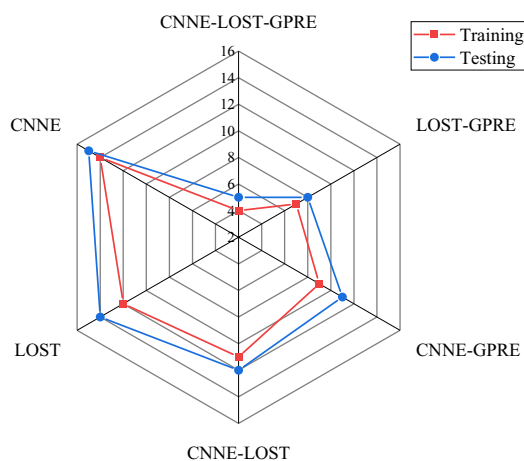
For irrigation purposes, it is necessary to evaluate the hydrochemical quality of groundwater. This section uses different indices to assess the water quality characteristics of the aquifer. Na<sup>+</sup> is one of the most important



(a) MAE

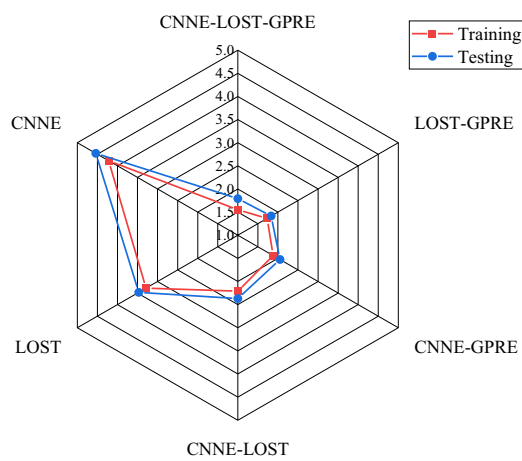


(b) NSE

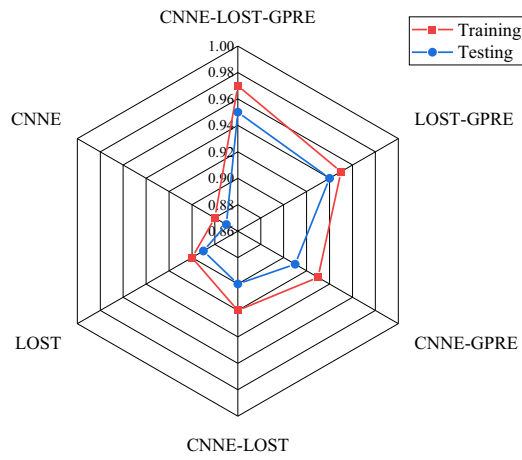


(c) PBIAS

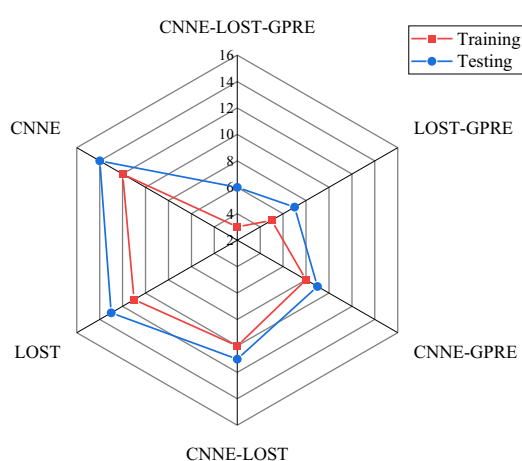
**Fig. 7** Radar plots of error indices for predicting EC, (a) MAE, (b) NSE, and (c) PBIAS



(a) MAE

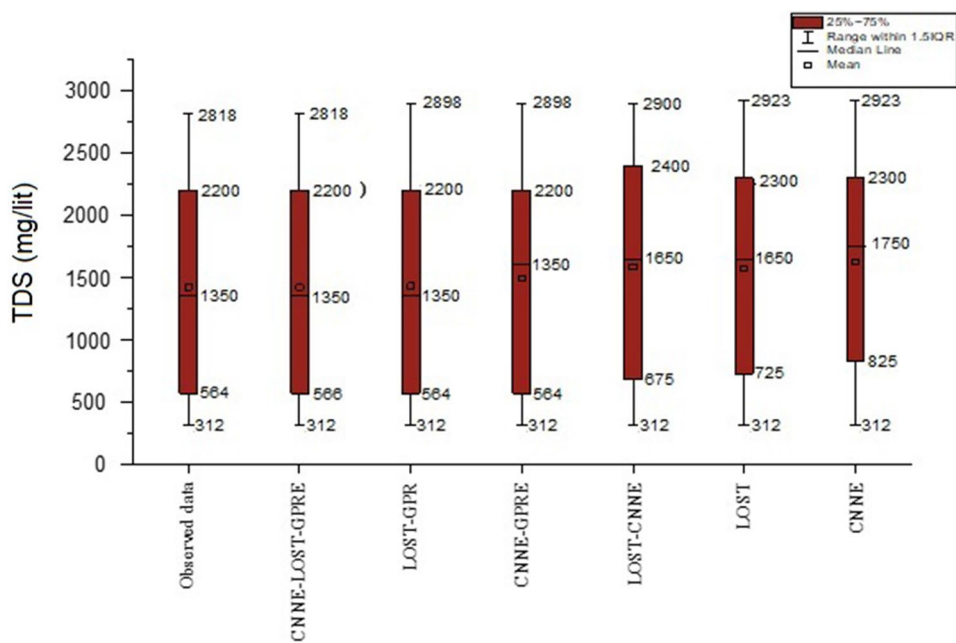


(b) NSE

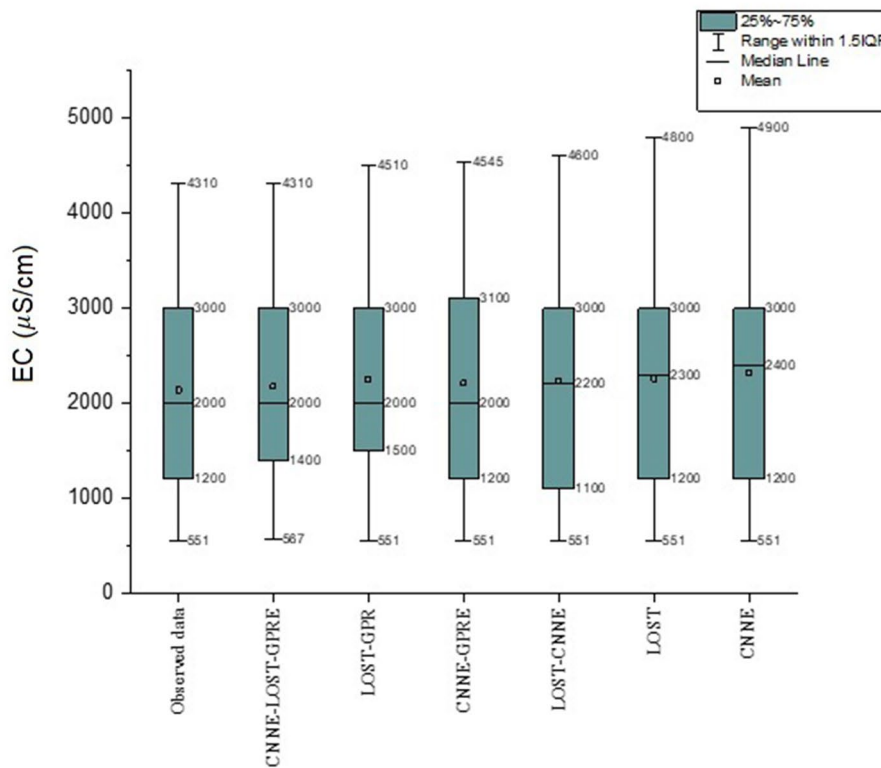


(c) PBIAS

**Fig. 8** The radar plots of error indicate predicting TDS, (a) MAE, (b) NSE, and (c) PBIAS

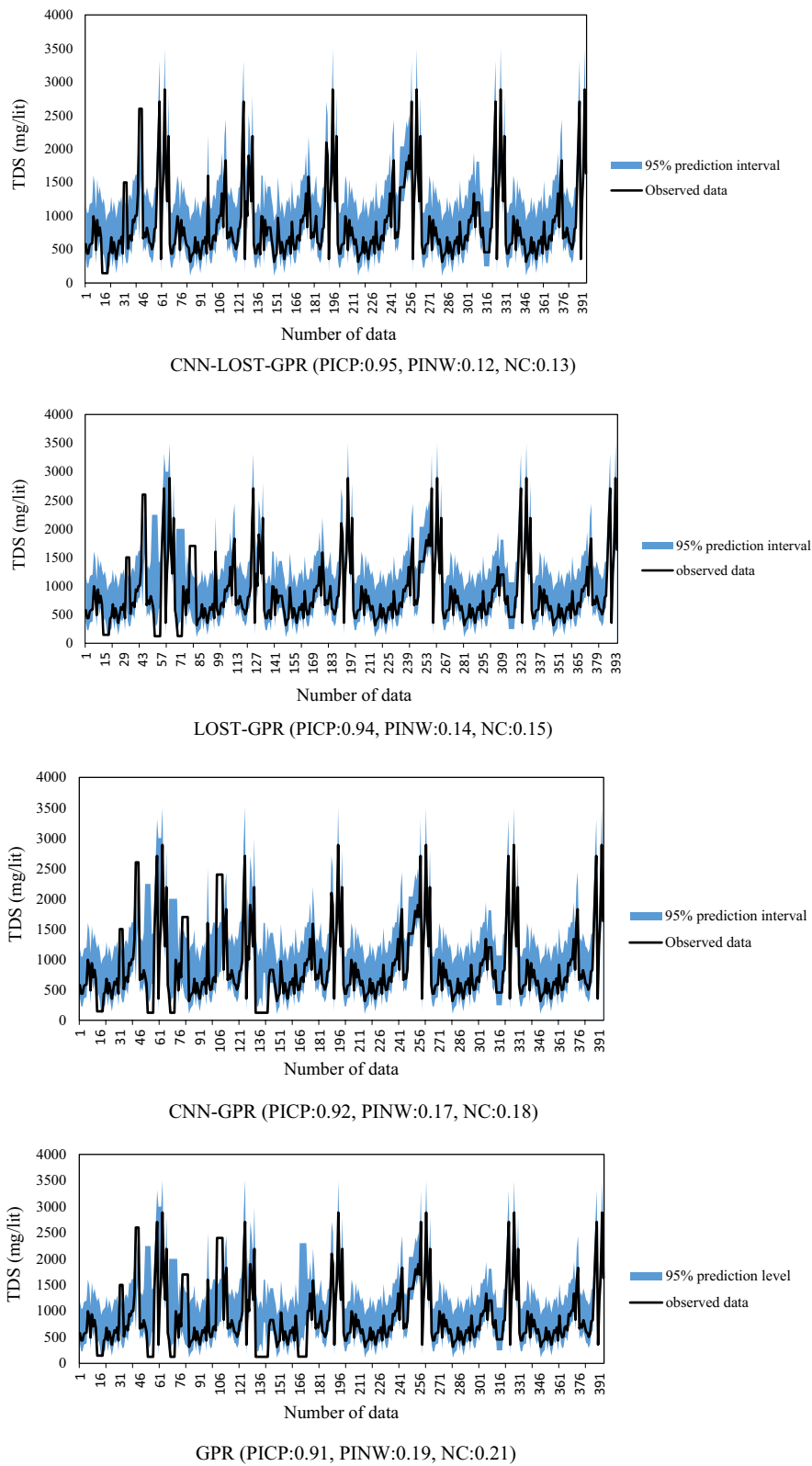


(a) TDS

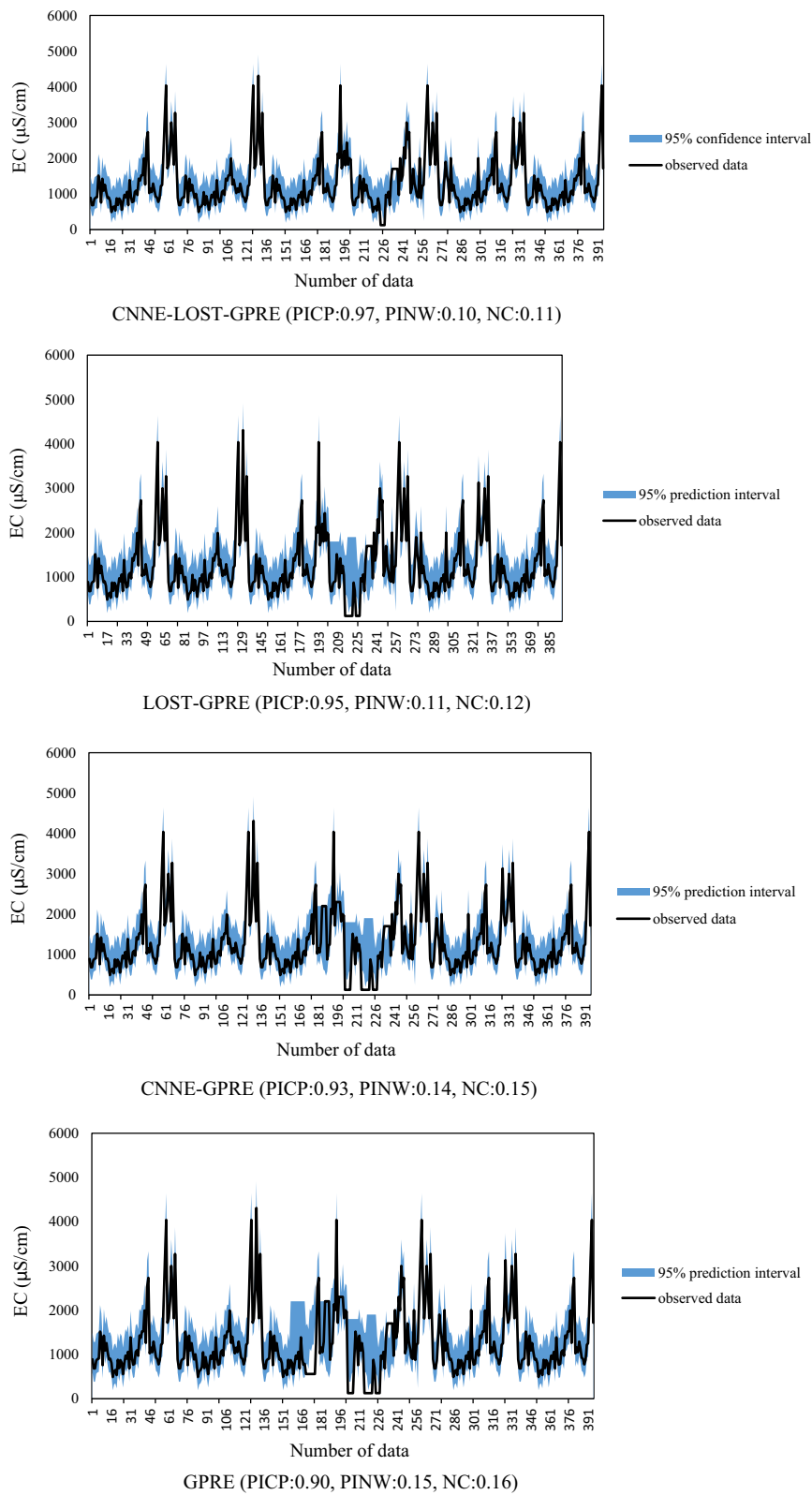


(b) EC

Fig. 9 Boxplots of models for comparison of the models, (a) TDS, and (b) EC



**Fig. 10.** 95% confidence interval for predicting TDS



**Fig. 11** The 95% prediction interval of TDS predictions



**Table 4** Summary of PICP, PINW, and NC results of 95% prediction interval for TDS and EC

| Parameters | Models         | PICP | PINW | NC   |
|------------|----------------|------|------|------|
| TDS        | CNN-LOST-GPR   | 0.95 | 0.12 | 0.13 |
|            | LOST-GPR       | 0.94 | 0.14 | 0.15 |
|            | CNN-GPR        | 0.92 | 0.17 | 0.18 |
|            | GPR            | 0.91 | 0.19 | 0.21 |
| EC         | CNNE-LOST-GPRE | 0.97 | 0.10 | 0.11 |
|            | LOST-GPRE      | 0.95 | 0.11 | 0.12 |
|            | CNNE-GPRE      | 0.93 | 0.14 | 0.15 |
|            | GPRE           | 0.90 | 0.15 | 0.16 |

**Table 5** The classification of water samples

| Parameter | Excellent | Good        | Doubtful   | Unsuitable |
|-----------|-----------|-------------|------------|------------|
| SRA       | –         | 45%         | 33%        | 22%        |
| Parameter | Excellent | Good        | Doubtful   | Unsuitable |
| EC        | –         | 10%         | 67%        | 23%        |
| Parameter | Suitable  |             | Unsuitable |            |
| MHR       | 22        |             | 78         |            |
| Parameter | Good      | Permissible | Doubtful   | Unsuitable |
| Sodium%   | 50%       | 20%         | 10%        | 10%        |
|           | Soft      | Moderate    | Hard       | Very hard  |
| TH        | –         | –           | 70         | 30         |

parameters for evaluating water quality. When sodium levels exceed the safe level, water permeability is reduced, and crops are damaged.

The classification of water samples is shown in Table 5.

- SRA

Based on SRA, 45, 33, and 22% of the water samples are good, doubtful, and unsuitable, respectively. If the SRA of water is high, it may cause the dispersion of soil colloids.

- MHR

Too much magnesium inhibits calcium absorption, and plant growth is reduced. 78 and 22% of samples are suitable and unsuitable based on the MHR parameter. Thus, water can adversely affect crop growth.

- EC

Higher EC inhibits nutrient uptake by increasing the osmotic pressure of the nutrient solutions. The health and yield of plants may be severely affected by lower EC. Based on EC values, 10, 67, and 23% of water samples are good, doubtful, and unsuitable, respectively.

- Sodium%

Crop yield is reduced when the sodium concentration exceeds the permissible limit. 50, 20, 10, and 10% of water samples were good, permissible, doubtful, and unsuitable.

- TH

Based on THE values, 70 and 30% of water samples were hard and unsuitable. Thus, THE values indicate the low quality of water samples.

Based on the comparison of the utilized and developed hybrid machine learning models, it shows that CNN-LOST-GPR outperformed other proposed models (LOST-GPRE, CNNE-GPRE, GPRE) in predicting TDS and EC. This study demonstrates that the CNNE-LOST-GPRE model is a reliable predictor of complex occurrences. As a result, the already developed hybrid model could be utilized by the private and public water sectors to estimate TDS and EC in coastal aquifers in order to improve water quality. While population and irrigation demand may increase in the future, water quantity and quality are poor. Hence, decision-makers must develop new policies and strategies for managing the basin’s water quality. In most cases, water table levels and subsidence are reduced, and water quality is improved through recharge basins. Brackish groundwater desalination is another widely used method in different world regions. Moreover, based on the PICP of the 95% prediction interval results for TDS, CNN-LOST-GPR outperformed LOST-GPR, CNN-GPR, and GPR with PICP of 0.95, 0.94, 0.91, and 0.91 respectively. Furthermore, based on the PICP of the 95% prediction interval results for EC, CNNE-LOST-GPRE outperformed LOST-GPRE, CNNE-GPRE, and GPRE with PICP of 0.97, 0.95, 0.93, and 0.90 respectively.

There are various advantages of the CNNE-LOST-GPRE hybrid model. For instance, CNN is able to capture both short-term and long-term dependency. LOST is able to intricate temporal dependency patterns. GPR could yield reasonable intervals for projected states, which is valuable for estimating uncertainty. Therefore, those three algorithms could attain a well performed accurate model. Besides, there are some limitations of the CNNE-LOST-GPRE hybrid machine learning model. For instance, CNN tends to be slow and training the data takes a long time. Furthermore, when the training data is limited or noisy, LSTM tends to overfit and lose generalization ability. Finally, GPR assumes a normal distribution, which is inappropriate for variables with only positive values.

## Conclusion

The study proposed a new hybrid model, CNN-LOST-GPR, to predict EC and TDS in the Qaemshahr costa aquifer. The new model predicts points and intervals simultaneously. CNN identifies features automatically. Using the GPR, intervals can be predicted. PH, Ca<sup>++</sup>, Mg<sup>++</sup>, Na<sup>+</sup>, K<sup>+</sup>, HCO<sub>3</sub><sup>-</sup>, SO<sub>4</sub><sup>-</sup>, and Cl<sup>-</sup> were used to predict EC and TDS. The RSOA was used for adjusting model parameters. The CNNE-LOST-GPRE was superior to other models. The testing PBIAS of the CNNE-LOST-GPRE, LOST-GPR, CNNE-GPRE, CNNE-LOST, LOST, and CNN models were 6, 7, 9, 11, 13, and 14 for predicting TDS. The training MAE of the CNN-LOST-GPR, LOST-GPRE, CNNE-GPRE, CNNE-LOST, LOST, and CNNE models were 1.67 mg/lit, 1.75 mg/lit, 1.9 mg/lit, 2.23 mg/lit, 3.24 mg/lit, and 4.25 mg/lit for predicting EC. In the modeling process, CNNE-LOST-GPRE provided lower uncertainty. Among the other models, LOST and CNNE had the lowest performance. Based on the results, CNNE-LOST-GPRE is a reliable model for extracting features and predicting outputs. The models help decision-makers when they encounter many features. SRA, EC, MHR, sodium percentage, and total hardness values indicated poor groundwater quality. In future research, CNNE-LOST-GPRE could be used to predict other characteristics of water quality. In addition, other optimization algorithms can also be investigated to improve the accuracy of the proposed hybrid model.

## Author contributions

ZJ: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft; SDL: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft; ME: Conceptualization, Data acquisition, Analysis and interpretation of data, Writing—original draft; ZSK: Conceptualization, Analysis and interpretation of data, Writing—original draft; ANA: Conceptualization, Investigation, Visualization; MS: Reviewing, Editing and Finalizing this paper, funding.; AES: Supervision, Writing—original draft.

## Funding

APC will be covered by National Water and Energy Center, United Arab Emirates University, United Arab Emirates.

## Availability of data and materials

Some data is available from the corresponding author upon request.

## Declarations

### Ethics approval and consent to participate

The authors undertake that this article has not been published in any other journal and that no plagiarism has occurred. The authors agree to participate in the journal.

### Consent for publication

The authors agree to publish it in the journal.

### Competing interests

The authors have no relevant financial or non-financial interests to disclose.

## Author details

<sup>1</sup>Department of Civil Engineering, University of Kashan, Kashan, Iran. <sup>2</sup>Scientific Research Center, Soran University, Soran, Erbil, Kurdistan Region, Iraq. <sup>3</sup>Civil Engineering Department, College of Engineering, Komar University of Science and Technology, Sulaimany, Kurdistan Region, Iraq. <sup>4</sup>Department of Water Engineering, Semnan University, Semnan, Iran. <sup>5</sup>Paleoclimate Dynamics Group, Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, 27570 Bremerhaven, Germany. <sup>6</sup>School of Engineering and Technology, Sunway University, Bandar Sunway, 47500 Petaling Jaya, Malaysia. <sup>7</sup>Civil and Environmental Eng. Dept., College of Engineering, United Arab Emirates University, 15551 Al Ain, United Arab Emirates. <sup>8</sup>Department of Civil Engineering, Faculty of Engineering, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia.

Received: 26 April 2023 Accepted: 23 January 2024

Published online: 30 January 2024

## References

- Ahmed AAM, Shah SMA (2017) Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. *J King Saud Univ Eng Sci*. <https://doi.org/10.1016/j.jksues.2015.02.001>
- Amanollahi J, Kaboodvandpour S, Majidi H (2017) Evaluating the accuracy of ANN and LR models to estimate the water quality in Zarivar International Wetland, Iran. *Nat Hazards*. <https://doi.org/10.1007/s11069-016-2641-1>
- Antony S, Dev VV, Kaliraj S, Ambili MS, Krishnan KA (2020) Seasonal variability of groundwater quality in coastal aquifers of Kavaratti Island, Lakshadweep Archipelago, India. *Groundw Sustain Dev*. <https://doi.org/10.1016/j.gsd.2020.100377>
- Azrou M, Mabrouki J, Fattah G, Guezaz A, Aziz F (2022) Machine learning algorithms for efficient water quality prediction. *Model Earth Syst Environ*. <https://doi.org/10.1007/s40808-021-01266-6>
- Banadkooki FB, Ehteram M, Panahi F, Sh. Sammen S, Othman FB, EL-Shafie A (2020) Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *J Hydrol*. <https://doi.org/10.1016/j.jhydrol.2020.124989>
- Barchi F, Parisi E, Urgese G, Ficarra E, Acquaviva A (2021) Exploration of convolutional neural network models for source code classification. *Eng Appl Artif Intell* 97:104075. <https://doi.org/10.1016/j.engappai.2020.104075>
- Barzegar R, Adamowski J, Moghaddam AA (2016) Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. *Stoch Environ Res Risk Assess*. <https://doi.org/10.1007/s00477-016-1213-y>
- Barzegar R, Asghari Moghaddam A, Adamowski J, Ozga-Zielinski B (2018) Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. *Stoch Env Res Risk Assess*. <https://doi.org/10.1007/s00477-017-1394-z>
- Charulatha G, Srinivasalu S, Uma Maheswari O, Venugopal T, Giridharan L (2017) Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arab J Geosci*. <https://doi.org/10.1007/s12517-017-2867-6>
- Chen Y, Song L, Liu Y, Yang L, Li D (2020) A review of the artificial neural network models for water quality prediction. *Appl Sci*. <https://doi.org/10.3390/app10175776>
- Chen W, Jiang M, Zhang WG, Chen Z (2021) A novel graph convolutional feature based convolutional neural network for stock trend prediction. *Info Sci* 556:67–94. <https://doi.org/10.1016/j.ins.2020.12.068>
- Dhaka VS et al (2021) A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. *Sensors*. <https://doi.org/10.3390/s21144749>
- Dhiman G, Garg M, Nagar A, Kumar V, Dehghani M (2021) A novel algorithm for global optimization: rat swarm optimizer. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-020-02580-0>
- Ghimire S, Yaseen ZM, Farooque AA, Deo RC, Zhang J, Tao X (2021) Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Sci Rep* 11(1):1–26. <https://doi.org/10.1038/s41598-021-96751-4>

15. Han D, Currell MJ (2022) Review of drivers and threats to coastal groundwater quality in China. *Sci Total Environ.* <https://doi.org/10.1016/j.scitotenv.2021.150913>
16. Jamei M, Ahmadianfar I, Olumegbon IA, Karbasi M, Asadi A (2021) On the assessment of specific heat capacity of nanofluids for solar energy applications: application of Gaussian process regression (GPR) approach. *J Energy Stor* 33:102067. <https://doi.org/10.1016/j.est.2020.102067>
17. Khadr M, Elshemy M (2017) Data-driven modeling for water quality prediction case study: the drains system associated with Manzala Lake, Egypt. *Ain Shams Eng J* 8(4):549–557. <https://doi.org/10.1016/j.asej.2016.08.004>
18. Khashei-Siuki A, Sarbazi M (2015) Evaluation of ANFIS, ANN, and geostatistical models to spatial distribution of groundwater quality (case study: Mashhad plain in Iran). *Arab J Geosci* 8(2):903–912. <https://doi.org/10.1007/s12517-013-1179-8>
19. Kouadri S, Elbeltagi A, Islam ARM, Kateb S (2021) Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Appl Water Sci* 11(12):1–20. <https://doi.org/10.1007/s13201-021-01528-9>
20. Kumari P, Toshiwal D (2021) Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Appl Energy* 295:117061. <https://doi.org/10.1016/j.apenergy.2021.117061>
21. Latif SD, Nor Azmi MS, Ahmed AN, Fai CM (2021) Application of artificial neural network for forecasting nitrate concentration as a water quality parameter: a case study of Feitsui Reservoir, Taiwan. *Int J Design Nat Ecodynam.* <https://doi.org/10.1828/jjdne.150505>
22. Latif SD, Birima A, Ahmed AN, Hatem DM, Al-Ansari N, Fai CM, El-Shafie A (2022) Development of prediction model for phosphate in reservoir water system based machine learning algorithms. *Ain Shams Eng J* 13:1. <https://doi.org/10.1016/j.asej.2021.06.009>
23. Liu Y, Li D, Wan S, Wang F, Dou W, Xu X, Qi L (2022) A long short-term memory-based model for greenhouse climate prediction. *Int J Intell Syst* 37(1):135–151. <https://doi.org/10.1002/int.22620>
24. Mattas C, Dimitraki L, Georgiou P, Venetsanou P (2021) Use of factor analysis (FA), artificial neural networks (ANNs), and multiple linear regression (MLR) for electrical conductivity prediction in aquifers in the gallikos river basin, northern Greece. *Hydrology* 8(3):127. <https://doi.org/10.3390/hydrology8030127>
25. Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA, Chan NW (2015) Prediction of water quality index in constructed wetlands using support vector machine. *Environ Sci Pollut Res* 22(8):6208–6219. <https://doi.org/10.1007/s11356-014-3806-7>
26. Najafabadipour A, Kamali G, Nezamabadi-pour H (2022) Application of artificial intelligence techniques for the determination of groundwater level using spatio–temporal parameters. *ACS Omega.* <https://doi.org/10.1021/acsomega.2c00536>
27. Najafabadipour A, Kamali G, Nezamabadi-pour H (2022) The innovative combination of time series analysis methods for the forecasting of groundwater fluctuations. *Water Resour* 49:283–291. <https://doi.org/10.1134/S0097807822020026>
28. Najah A, Teo FY, Chow MF et al (2021) Surface water quality status and prediction during movement control operation order under COVID-19 pandemic: case studies in Malaysia. *Int J Environ Sci Technol* 18:1009–1018. <https://doi.org/10.1007/s13762-021-03139-y>
29. Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H, Khaleel Ibrahim R, Ming Fai C, Shabbir Hossain M, Ehteram M, Elshafie A (2019) Machine learning methods for better water quality prediction. *J Hydrol.* <https://doi.org/10.1016/j.jhydrol.2019.124084>
30. Ouma YO, Cheruyot R, Wachera AN (2022) Rainfall and runoff time-series trend analysis using LOST recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of Nzoia hydrologic basin. *Complex Intell Syst* 8(1):213–236. <https://doi.org/10.1007/s40747-021-00365-2>
31. Ravansalar M, Rajae T (2015) Evaluation of wavelet performance via an ANN-based electrical conductivity prediction model. *Environ Monit Assess* 187(6):1–16. <https://doi.org/10.1007/s10661-015-4590-7>
32. Reynolds LB (2000) Nutrient solution bicarbonate concentration effects on growth media pH, electrical conductivity, and on flue-cured tobacco yield and quality for transplants produced in a heated greenhouse in a direct-seeded float system. *Tob Sci.* <https://doi.org/10.3381/0082-4623-44.1.27>
33. Salami ES, Salari M, Ehteshami M, Bidokhti NT, Ghadimi H (2016) Application of artificial neural networks and mathematical modeling for the prediction of water quality variables (case study: southwest of Iran). *Desalin Water Treat* 57(56):27073–27084. <https://doi.org/10.1080/19443994.2016.1167624>
34. Sarvamangala DR, Kulkarni RV (2022) Convolutional neural networks in medical image understanding: a survey. *Evol Intel.* <https://doi.org/10.1007/s12065-020-00540-3>
35. Senthilkumar S, Gowtham B, Sundararajan M, Chidambaram S, Lawrence JF, Prasanna MV (2018) Impact of landuse on the groundwater quality along coastal aquifer of Thiruvallur district, South India. *Sustain Water Resour Manage* 4(4):849–873. <https://doi.org/10.1007/s40899-017-0180-x>
36. Wan X, Li X, Wang X, Yi X, Zhao Y, He X, Huang M (2022) Water quality prediction model using Gaussian process regression based on deep learning for carbon neutrality in papermaking wastewater treatment system. *Environ Res* 211:112942. <https://doi.org/10.1016/j.envres.2022.112942>
37. Wang H, Zhang YM, Mao JX (2022) Sparse Gaussian process regression for multi-step ahead forecasting of wind gusts combining numerical weather predictions and on-site measurements. *J Wind Eng Ind Aerodyn* 220:104873. <https://doi.org/10.1016/j.jweia.2021.104873>
38. Xu Y, Hu C, Wu Q, Jian S, Li Z, Chen Y, Wang S (2022) Research on particle swarm optimization in LOST neural networks for rainfall-runoff simulation. *J Hydrol* 608:127553. <https://doi.org/10.1016/j.jhydrol.2022.127553>
39. Yan R, Liao J, Yang J, Sun W, Nong M, Li F (2021) Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LOST, CNN-LOST, and spatiotemporal clustering. *Expert Syst Appl* 169:114513. <https://doi.org/10.1016/j.eswa.2020.114513>
40. Zhang L, Zou Z, Shan W (2017) Development of a method for comprehensive water quality forecasting and its application in Miyun reservoir of Beijing, China. *J Environ Sci* 56:240–246. <https://doi.org/10.1016/j.jes.2016.07.017>
41. Zhang Z, Ye L, Qin H, Liu Y, Wang C, Yu X, Li J (2019) Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression. *Appl Energy* 247:270–284. <https://doi.org/10.1016/j.apenergy.2019.04.047>
42. Zhu S, Heddam S (2020) Prediction of dissolved oxygen in urban rivers at the three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). *Water Quality Res J* 55(1):106–118. <https://doi.org/10.2166/wqrj.2019.053>
43. Ziyad Sami BF, Latif SD, Ahmed AN et al (2022) Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan. *Sci Rep* 12:3649. <https://doi.org/10.1038/s41598-022-06969-z>

## Publisher's Note

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