Water Supply

Water Supply Vol 22 No 5, 5237 doi: 10.2166/ws.2022.165

A hybrid artificial neural network: An optimization-based framework for smart groundwater governance

Asmae El Mezouari *, Abdelaziz El Fazziki and Mohammed Sadgal

Computer Systems, Cadi Ayyad University, B. P549, Av. Abdelkarim, Elkhattabi Guéliz, Marrakech, Morocco
*Corresponding author. E-mail: asmae.elmezouari@ced.uca.ma

AEM, 0000-0001-6096-5689

ABSTRACT

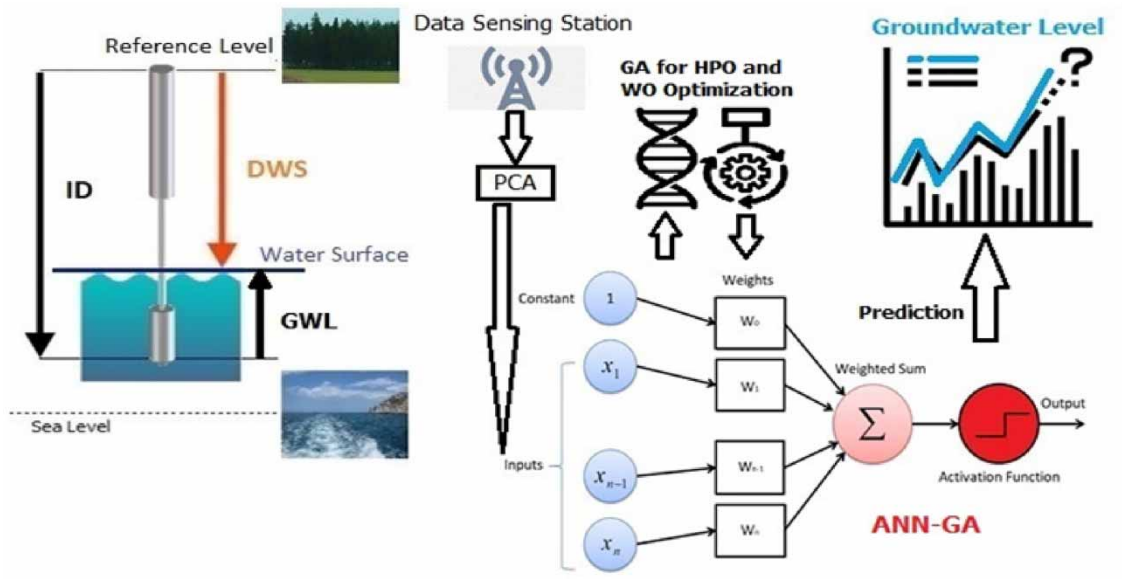
Given the growing scarcity and strong demand for water resources, the sustainability of water resource management requires an urgent policy of measures to ensure the rational use of these resources. The heterogeneous properties of groundwater systems are related to the dynamic temporal-spatial patterns that cause great difficulty in quantifying their complex processes, while good regional groundwater level forecasts are completely required for managing water resources to guarantee suitable support of water demands within any area. Water managers and farmers need intelligent groundwater and irrigation planning systems and effective mechanisms to benefit from the scientific and technological revolution, particularly the artificial intelligence engines, to enhance the water support in their water use planning practices. Therefore, this work aims to improve the groundwater level prediction based on the previous measures for better planning of hydraulic resource use. For this concern, the suggested method starts with data-preprocessing using the Principal Component Analysis method. Next, we validated the effectiveness of the hybrid artificial neural network, combined with an extended genetic algorithm for the hyperparameters and weight optimization, in predicting the groundwater levels in a selected monitoring well in California. The evaluation results have demonstrated the performance of the optimized ANN-GA model.

Key words: artificial neural networks, genetic algorithm, groundwater level prediction, knowledge extraction, principal component analysis

HIGHLIGHTS

- Smart Groundwater Governance Framework.
- Pertinent Data Extraction with Principal Component Analysis.
- The Artificial Neural Networks' hyper-parameters optimization.
- The Artificial Neural Networks' weights optimization.
- Improving the groundwater levels prediction is primordial to sustainable water management.

GRAPHICAL ABSTRACT



1. INTRODUCTION

In the last decades, groundwater has become the most important source of the freshwater serving of global water demands in the world. Hence, the availability of groundwater is the most crucial factor impacting socioeconomic development. Adding to that, given the climate changes nowadays, the impact of non-integrated water resources management is becoming a global problem all over the world, especially in countries characterized by overpopulation or intensive agricultural activities (Lall *et al.* 2020). In addition, domestic water needs, industrial and irrigation requirements increased the demand for groundwater. Therefore, desalination constitutes a primordial pillar for serving the increasing water demand for drinking, irrigation, and industries. However, desalination may deeply threaten the natural environment, according to Panagopoulos (2021a). For this reason, many studies are directed to develop safe desalination treatments (Panagopoulos 2021b, 2021c). Moreover, because of the important decrease in freshwater, the critical groundwater level monitoring allows consistent policies to decrease some constraints related to sustainable water management, like the water loss of pumping in wells dedicated to domestic water supply, aquifer compaction, and land surface subsidence (Guzy & Malinowska 2020). Likewise, many pieces of research have been conducted to study and detect the internal relationships between the hydrological and soil dynamics parameters or to model and evaluate the quality of water based on various Artificial Neural Networks, Adaptive Neuro-Fuzzy Inference Systems, and curve fitting (Alrashed *et al.* 2018a; Bahrami *et al.* 2019; Ghasemi *et al.* 2019; Moradikazerouni *et al.* 2019; Safaei *et al.* 2019; Khosravi *et al.* 2021). However, systems allowing monitoring of groundwater levels, groundwater quality, and land subsidence are too expensive (Edwards & Guilfoos 2021). Consequently, groundwater level assessment is usually taken as a priority over controlling groundwater quality and land subsidence. In some previous research works, most of the hydrological studies rely on physical and conceptual models to detect and describe hydrological parameters, resulting from physical processes in the ground (Alrashed *et al.* 2018b; Karimipour *et al.* 2018; Bagherzadeh *et al.* 2019; Karimipour *et al.* 2019; Peng *et al.* 2020; Giwa *et al.* 2021). However, some practical limitations related to extracting and gathering the data still exist. Furthermore, a new ensemble modeling framework that relies on spectral analysis, machine learning, and uncertainty analysis has been deployed by Sahoo *et al.* (2017) to study climate change, surface water flows, and irrigation activity relationships for a better understanding and prediction of the groundwater level change. The approach was applied to two aquifer systems that support production in agriculture in the United States, starting from selecting input data sets based on mutual information, genetic algorithms (GAs), and lag analysis, to using the selected data sets in a Multilayer Perceptron Artificial Neural Network architecture to simulate seasonal groundwater level change. The results show that the huge amounts of agricultural and hydrogeological data required for building and calibrating models for water resource management systems increase the complexity of their integration and sustainability.

Recently, a multitude of researchers have conducted comparative studies on various machine learning algorithms in groundwater prediction, including Artificial Neural Networks (ANN) such as in Wen *et al.* (2017) where the results of the comparison show that the wavelet analysis combined with the Artificial Neural Network (WA-ANN) outperforms the basic ANN in terms of accuracy using different data samples, including groundwater level and climatic data. Furthermore, a study by Wunsch *et al.* (2021) focused on comparing different ANN architectures, such as Conventional Recurrent Artificial Neural Networks, especially the non-linear autoregressive networks with exogenous input (NARX) and famous deep learning ANN models like long short-term memory (LSTM) and convolutional neural networks (CNNs); in terms of efficiency and applying different training methods for an accurate prediction of groundwater level in the long term. The results showed that the NARX superficial neural networks might outperform the Deep Learning algorithms, specifically in dealing with limited training data, where they can exceed both LSTMs and CNNs. However, LSTMs and CNNs might compute more precisely with a larger dataset. In another context, regarding the importance of applying reliable and accurate models for the prediction of groundwater level beneath evolving climatic circumstances, a recent study in Müller *et al.* (2021) introduces the primordial hyperparameter optimization in boosting the performance of machine learning models. In this experiment, researchers conducted a comparative study to evaluate the performance of three methods for hyperparameter selection, especially a random sampling method, and two surrogate model-based algorithms. They applied these methods to four modern deep learning (deep ANN) computational models, and the empirical results generated from all trained models show that the optimization of the hyperparameters gives reasonable and accurate performance. Ordinarily, the basic backpropagation artificial neural networks (BP-ANN) are applied with random hyperparameters, and weight initialization resulting from the symmetry breaking. Nevertheless, the random initialization of the hyperparameters and weights for training the neural networks may lead to trapping in local minima and slow convergence. Thus, the basic ANN models used in previous studies are usually not enough for the satisfaction of the desired solution towards a reliable predictive model for anticipating the future change of available water to act more accurately. Hence, data extraction and optimization-based GAs could be used as the magic solution to overcome these issues and improve the performance and efficiency of the ANN algorithms. Therefore, this work presents an interesting data extraction and a hybrid machine learning optimization based predictive approach that enhances the groundwater level change prediction accuracy required for sustainable groundwater consumption planning; by improving the training of the ANN model and enhancing its efficiency, precision, and robustness. This method relies on artificial intelligence tools, especially the ANNs, combined with the GA for the hyperparameters and weight optimization, hydrological cycle knowledge, and data mining tools. The rest of this paper is organized as follows. Firstly, section 2 exposes the proposed framework's architecture, the groundwater level, and depth measurement methods, and the different data processing and evaluation methods used in this study. Then, section 3 depicts the results of the proposed approach with a case study of groundwater level prediction in a well-monitoring station in California, US. Finally, this work ends with a concise conclusion and future perspectives.

2. MATERIALS AND METHODS

2.1. The suggested framework

The proposed framework, presented in Figure 1, intends to estimate and predict the groundwater surface level based on the depth to water surface change and other groundwater monitoring parameters that have been made by relying on hysterical groundwater level sensing in the sited wells to help in scheduling and pumping groundwater for industrial and irrigation use. The main goal of this component is therefore to facilitate the task for the water resource manager by ensuring an automatic estimation of the groundwater-surface level by predicting the water surface elevation based on the depth to water surface change and other relevant features extracted for selecting the appropriate predictive model, standing on the advances in data sensing, machine learning, and hyperparameter optimization tools. The principal components present in this framework are explained as follows: A data acquisition module that provides data on the wells, environment, and plots where the groundwater level data and reports are measured by direct manipulation in the well or sensed and saved automatically using other components like smart sensors, controllers, and storage devices.

Figure 2 is a data processing flux that employs various ANNs implemented in anaconda-python; particularly the basic feed-forward ANN, and two ANNs combined with the GA for hyperparameters and weights optimization; to predict the groundwater-surface level based on the depth to water surface change (DWS) retrieved from the sensed groundwater

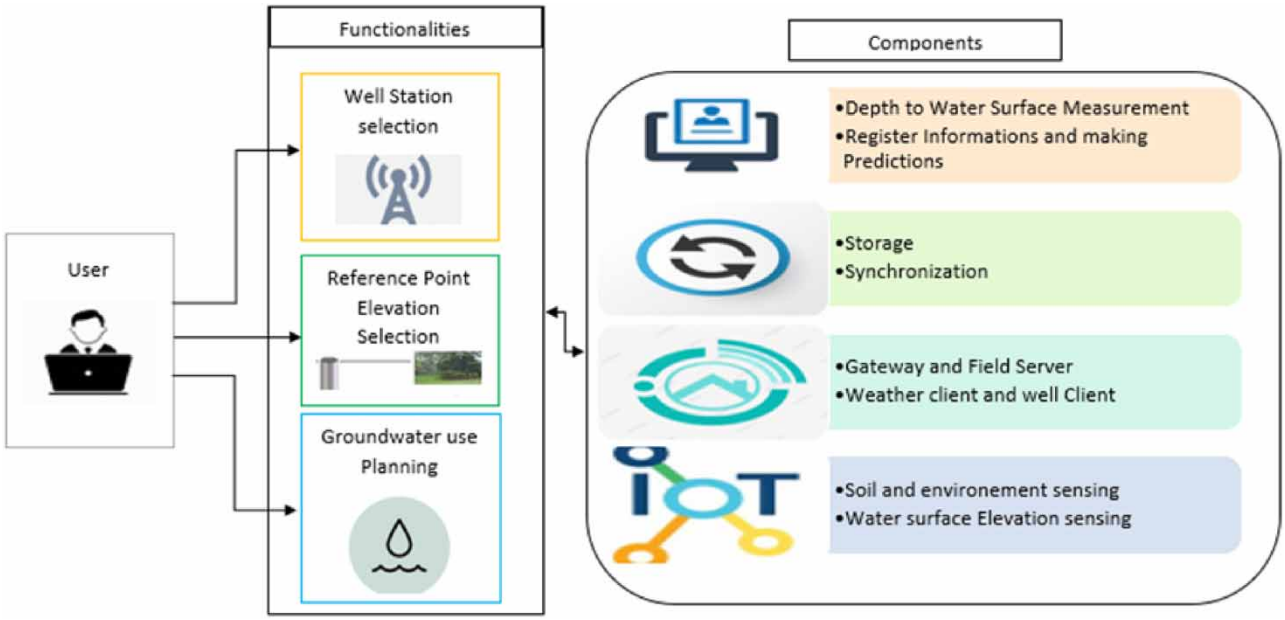


Figure 1 | The Proposed Framework' Architecture.

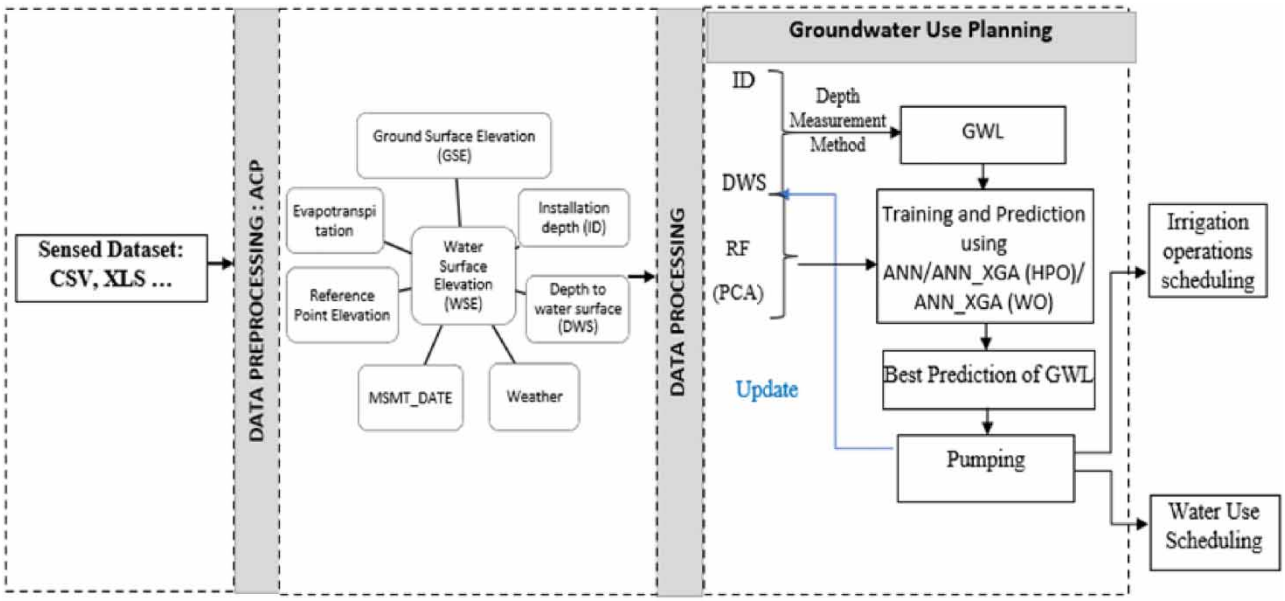


Figure 2 | Data processing flux.

depths measurements, the configured installation depth (ID), and the relevant features extracted from the PCA; to model future fluctuations in the water surface level regarding the water stress and recharge in the monitoring well. The data extraction aims to select the relevant features suitable for the learning process based on the Principal Component Analysis method. So, the studied approaches are compared in terms of efficiency based on the evaluation metrics described in section 2.6 for selecting the best estimation model that would be involved in groundwater level change prediction and to study the impact of the integrated extended GA used for optimizing both the initial hyperparameters and weights of the initial ANN model in terms of precision.

In this paper, we included a decision-making process in Figure 3 for groundwater monitoring and planning, which was designed based on the depth measurement method, surface water elevation sensing, well station configuration, and the scheduling management model. The irrigation and water use plan can be made based on the availability of groundwater depending on the estimated water surface elevation by predicting the groundwater level change of the aquifer or the land surface relying on historical and forecast weather data. This integrated process performs different groundwater level predictions based on the monitoring well station and the reference point configuration. The optimization of water use and the historical experience allow the user to personalize and schedule groundwater use. The principal function of this process includes a decision-making aid for managing the groundwater in the different phases of exploitation and aligning the groundwater use with climatic changes, taking into account hydrological dynamics on the land surface. A case study of groundwater management will be carried out to show the benefits of this process and its practical improvements.

2.2. Groundwater level and depth measurement method

Groundwater is defined as the water that has infiltrated into the surface of the ground and formed aquifers. The groundwater level is generally observed through appropriate sensing instruments recognized as monitoring wells. Monitoring wells are wells including a small-bore under the ground, employed for groundwater level monitoring and water quality examination. Water-level control is a major way to understand groundwater changes in basins caused by recharge (precipitation, irrigation return, seepage from streams, etc.) and discharge (groundwater pumping, seepage to streams, etc.), to determine directions of groundwater movement and trends in groundwater storage, and to evaluate progress toward sustainable water resource management. Moreover, groundwater level monitoring concerns continuous or periodic measures depending on the user configuration. Continuous measuring is done using automatic water-level sensors that are involved in monitoring measurements in wells. Consequently, it provides a high-resolution record of water-level fluctuations that can accurately identify the effects of various stresses on the aquifer system and provide measurements of maximum and minimum water levels in aquifers (Taylor & Alley 2001). In addition, groundwater measurements are often taken using a conductivity switch that is grounded into a borehole through a flat steel or plastic cable and emits an acoustic signal when it touches the surface water. Thus, the depth from the surface to the groundwater can be simply measured, but it requires a manual application. Nowadays, automatic groundwater level measurements can be taken through KELLER data loggers that collect the data autonomously using a level sensor, connected to a microprocessor circuit with a battery and a storage device. The user can set the periods of measurements to be performed and stored at each time by the logger. This sleep mode saves the battery

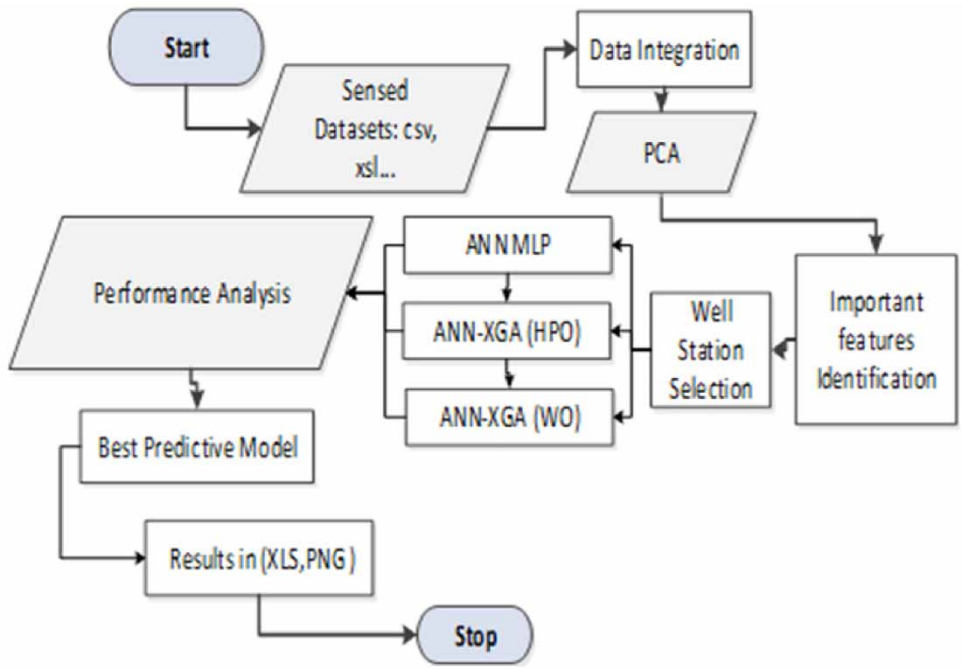


Figure 3 | The predictive process.

for up to ten years. Data is produced and delivered via a USB converter and Logger 5 software from KELLER. The groundwater level (GWL) is measured by the data loggers via the membrane in the submersible transmitter. However, the largest number of geohydrologists remain engaged in measuring the distance from the top of the borehole to the actual water surface level in the borehole (DWS), as shown in the graphical abstract. The depth to the water surface level is then calculated based on the water column using the formula as follows:

$$DWS = ID - GWL \quad (1)$$

where ID is the total installation depth that can be set as a passive parameter in the data logger, and GWL represents the measured groundwater column.

Hence, continuous monitoring may not only be the most convenient way to monitor fluctuations in groundwater levels during droughts and crucial stages when hydraulic pressures may change at very fast rates, but also when real-time measurements are required for water decision-making. Moreover, near-continuous data collection can be performed by using telecommunication or radio transmitter devices at the monitoring site.

2.3. Artificial neural networks

ANNs are the most popular and powerful subfield of machine learning, including algorithms inspired by brain processing to help in decision-making (Schmidhuber 2015). A single-layer neural network that exports a single output has been denoted as the perceptron, according to Gupta (2013). Recently, many subjects have proved the efficiency of ANN algorithms for predicting the future state of the studied features in various fields.

2.4. Evolutionary algorithms (EAs)

Hyperparameter tuning for machine learning processing is a challenging issue. Sometimes the obtained output may not be accurate due to the bad initialization of the parameters' values instead of resulting from noisy data or a weak model. In fact, the optimization process is required to find the optimal value for each hyperparameter, maximizing the performance.

Hence, according to Eiben & Smith (2003) many operation research (OR) researchers advise several optimization techniques such as evolutionary algorithms (EAs) for hyperparameter optimization. The main aspects of these optimization techniques are:

- Multimodal Optimization
- Multi-objective Optimization
- Constrained Optimization
- Combinatorial Optimization

EAs are distinguished from traditional algorithms by their dynamic aspects related to their ability to evolve.
The main characteristics of evolutionary algorithms can be summarized as:

- Population-Based: EAs intend to optimize any process where the initial solutions lead to bad results. This set of initial solutions from which better solutions are generated is called the population.
- Fitness-Oriented: the criteria measure that distinguishes each solution from another one is the fitness value, which is linked to each solution and calculated from a fitness function. The fitness value indicates how reliable the solution is.
- Variation-Driven: the updating operation of the current solutions happens whenever the current solutions do not satisfy the fitness function calculated. Hence the individual solutions are to evolve and undergo some variations to create new solutions that could meet the fitness criteria.

2.5. Extended genetic algorithm for artificial neural networks optimization

The GA belongs to evolutionary algorithms and relies on random changes to evolve the current solutions to find suitable solutions. The GA approach is founded on Darwin's theory of evolution, including a gradient descent by way of applying a slow, gradual process to make slight and slow changes in each generation until reaching the best solution. Each individual generated from the population is characterized by some properties represented by chromosomes (encoded properties) which can be adjusted (Do Crossover) or mutated (Do Mutation) or both, leading to a new generation of the

population being better in terms of fitness. According to Gad (2018), the GA disposes of several operations performed for optimization.

We have extended the GA for hyperparameters optimization described in Rowe & Colbourn (2003) by integrating a self-learning step (training) with a small subset of data in Figure 4, to avoid the problem of exponential complexity due to data training over many generations, and to improve the irritated weights in the fitness function to mimic the ability of a human to evolve its aspects before transmitting them to his children in the future generation. From the same perspective, we have integrated the self-learning step in Figure 5 into the GA for weight optimization proposed by Gad (2018). In the end, the selected solution could be trained with a choice between all records or a data subset depending on the dimension, the volume, the value of the data, and the power of the processing environment available. The main idea behind this proposition is the fact that training the ANN using the genetic algorithm with a subset of data optimizes the processing time and can overcome the problem of exponential complexity due to data training over many generations, which enhances the performance. While a past training of the selected solution might improve the efficiency of the model related to the quantity and quality of the data available, whether it is noisy or not, and related to the processing environment, whether it is powerful or not. Here the GA plays the role of initializing optimized hyperparameters and weights instead of random initialization.

2.6. Performance measures

To evaluate the performance of the chosen predictive models, we adopted the following equations, the root mean square error (RMSE), the mean square error, the mean absolute error (MAE), and the r-squared accuracy:

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{n} \frac{(yi - \overline{y_i})^2}{n}} \tag{2}$$

where n is the number of records, is the actual prediction of instance i, and $\overline{y_i}$ is the suitable output. It represents a type of general error measure. The precision is higher when the value of this error is lower. Perceive that the RMSE is estimated based on a similar scope as the resulting variable. In this matter, easy comparison between the RMSE of the predictive methods is sufficient to assess their performance when the output parameter is similar for all the predictive models.

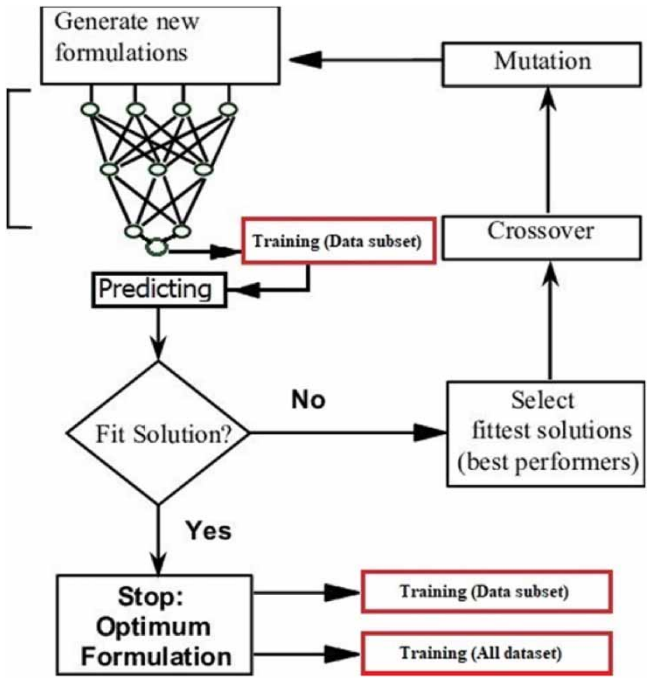


Figure 4 | The extended XGA structure for ANN' hyperparameters optimization.

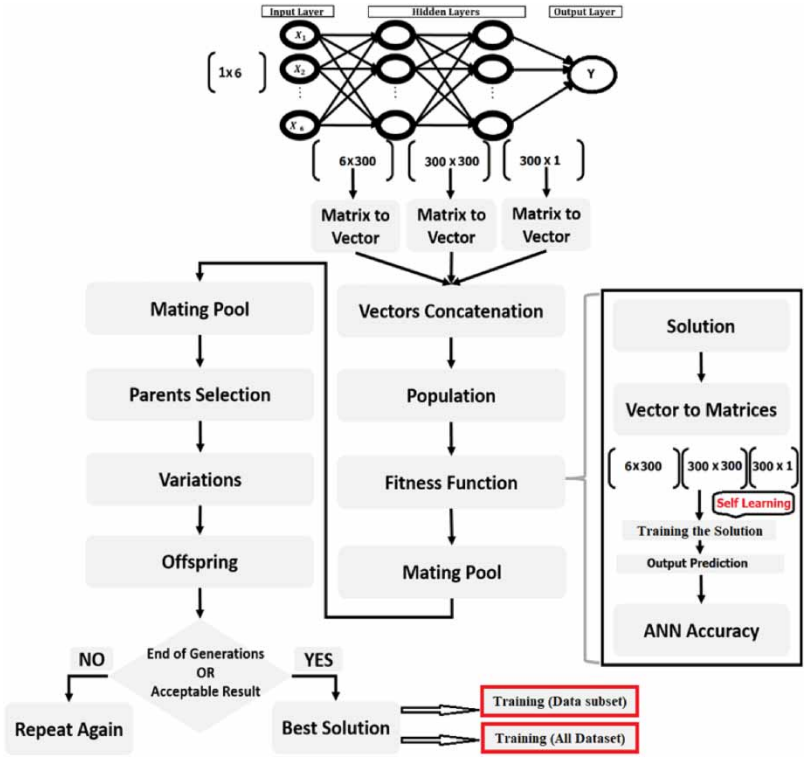


Figure 5 | The extended XGA Structure for ANN weighs optimization.

MAE is a measure that compares the errors between predicted and observed values describing the same phenomenon. The MAE is measured as:

$$\mathrm{MAE} = \sum_{i=1}^{n} \frac{|yi - xi|}{n} \tag{3}$$

R_2 evaluates how much the data is near the adjusted regression course. It may describe both determination and multiple determination for both single and multiple regression as presented in the formula:

$$R_2 \text{ accuracy} = \frac{\text{Explained variation}}{\text{Total variation}} \tag{4}$$

The R_2 score is a measure of percentage varying between 0 and 100:

- 0 expresses that the interpreted variability in the resulting data predicted by the model around its average is null.
- 100 means that the predictive model displays properly all the interpretations of variability in predicted data around its average.

2.7. Description of output and input variables

Concerning the available training data, the determination of training inputs was performed according to knowledge used by the water manager throughout the hydrological process and the groundwater availability estimation method, such as climatic data, water level, and the soil water state (depth reference point).

Though additional factors may influence the prediction of groundwater level (such as prediction time, potential suspension of irrigation in the area, the internal hydrological processes, etc.), these determinants are sometimes not taken into consideration by the experts. This represents a limitation of this strategy.

Thus, picking a suitable collection of factors is crucial for providing efficient predictive models. For this reason, the Principal Component Analysis (PCA) method will be applied to select the most important parameters.

2.8. Principal component analysis approach

PCA is a systematic approach used in multivariate statistics, that aims to convert some correlated variables into new irrelative variables. These created variables are designated 'principal components', the axes they define as 'main axes', and the correlated linear forms are 'main factors'. It affords an optimization of the set of variables by reducing information redundancy and improving relevancy (Guerrien 2003).

According to Ringnér (2008) and Jolliffe (2002), the main purposes attempted by PCA are:

- The 'optimal' graphic illustration of individuals (lines), decreases the deformations of the point cloud, in a subspace E of dimension $q(q<p)$.
- The graphical description of the variables in a subspace F by describing well the initial connections between these variables.
- The dimension decrease (compression), or approximation of X by an array of rank $q(q<p)$.

2.9. Case study: ground water level changes prediction

Due to the unavailability of data, we based our study on continuous groundwater level measurements in a dataset including daily and continuous time-series data collected from automatic recording instruments installed at many well sites in California and performed by the Department of Water Resources (CDWR 2019). The intervals between readings vary between 15 minutes and 1 hour. Some of the measures are delivered to the California Data Exchange Center. Nevertheless, most of the monitoring sites are controlled once every month or two, whenever measures are off-loaded from the data recorders and later finalized and published. The monitoring wells included in this dataset are located in Glenn, Butte, Glenn, Colusa, Modoc, Mendocino, Joaquin, Sacramento, San, Solano, Shasta, Solano, Tehama, Siskiyou, Tehama, Sutter, Yuba, and Yolo Counties. This dataset contains (1,048,576) records and includes about 12 variables necessary for the groundwater level estimation as shown in Table 1 and 2. The majority of the recorded data has a quality code of 1 or 2, which means, according to the definitions in Table 3, that the data is good for the study.

Table 1 | Dictionary of the dataset

Column	Type	Label	Description
STATION	text	Station	Unique Station Identifier (database key). For most stations, this is the State Well Number
MSMT_DATE	timestamp	Water Level Measurement Date (PST)	Date/Time (in PST) when the groundwater level measurement was collected
WLM_RPE	numeric	RPE for a specific water level measurement record	Reference Point Elevation used to collect the groundwater level measurement
WLM_RPE_QC	numeric	WLM_RPE Quality Code	Quality Code for WLM_RPE measurement
WLM_GSE	numeric	GSE for a specific water level measurement record	Ground Surface Elevation at the well site
WLM_GSE_QC	numeric	WLM_GSE Quality Code	Quality Code for WLM_GSE measurement
RPE_WSE	numeric	RPE to WSE	Depth to the water surface in feet below the reference point
RPE_WSE_QC	numeric	RPE to WSE Quality Code	Quality Code for RPE_WSE measurement
GSE_WSE	numeric	GSE to WSE	Depth below ground surface or the Distance from the ground surface to the water surface in feet
GSE_WSE_QC	numeric	GSE to WSE Quality Code	Quality Code for GSE_WSE measurement
WSE	numeric	WS Elevation	Water Surface Elevation in feet above Mean Sea Level (NAVD88)
WSE_QC	numeric	WS Elevation Quality Code	Quality Code for WSE measurement

Table 2 | Dataset structure

STATION	MSMT_DATE	WLM_RPE	WLM_RPE_QC	…	GSE_WSE	GSE_WSE_QC	WSE	WSE_QC
01N04E36Q001M	4/30/2005	9.1	1	…	15.154	1	−8.254	1
01N04E36Q001M	5/1/2005	9.1	1	…	15.148	1	−8.248	1
01N04E36Q001M	5/2/2005	9.1	1	…	15.143	1	−8.243	1
01N04E36Q001M	5/3/2005	9.1	1	…	15.158	1	−8.258	1
01N04E36Q001M	5/4/2005	9.1	1	…	15.154	1	−8.254	1
01N04E36Q001M	5/5/2005	9.1	1	…	15.119	1	−8.219	1
01N04E36Q001M	5/6/2005	9.1	1	…	15.114	1	−8.214	1
01N04E36Q001M	5/7/2005	9.1	1	…	15.122	1	−8.222	1
…	…	…	…	…	…	…	…	…

The groundwater level represents the major source of information about variations in groundwater storage and change in a basin, and how these are influenced by several forms of recharge (precipitation, infiltration from streams, irrigation return) and discharge (drainage to streams, groundwater use).

In the present case study, we focus on predicting the groundwater level of 15 months ahead of the monitoring well station '12N02E21Q003M' between 1/1/2021 and 9/29/2021, containing almost 271 days of the ANN models (basic and optimized using the extended genetic algorithms, XGA). Firstly, we analyzed the main components of the 12 parameters in the dataset presented in Table 1 to select the relevant set of parameters (seven relevant features in Table 4). Later, we performed predictions by applying supervised machine learning to the dataset in Python, using the Anaconda environment; over the records collected and estimated daily by experts at the USDA-Agricultural Research Service. Finally, we evaluated these models using several performance measures.

3. RESULTS AND DISCUSSION

In the current work, we intend to evaluate the impact of integrating hyperparameter and weight optimization on the prediction performance using an extended GA for optimizing the trained model's hyperparameters and weights; and compare it to the basic ANN(MLP) model. To this aim, we tried to apply two extended optimizers based on the GA, which is one of the simplest random-based EAs, to the ANN (multilayer perceptron). The first one is used for the hyperparameter optimization to

Table 3 | Quality codes

Quality Code	Description	Label	Description
1	Good data	201	Data not recorded
10	Good measurement	255	No data exists
104bb	Records estimated	40	Fair measurement
120	Poor measurement	50	Unknown measurement quality
130	Estimate	60	Above rating – extrapolated above 2x highest measurement; unreliable extrapolation
15	Provisional measurement	70	Estimated data
150	Rating table extrapolated due to inadequate gauging information	71	Manual reading
151	Data missing	76	Reliable interpolation
170	Unreliable data	85	Flooded
2	Good quality edited data	–	–

Table 4 | Dictionary of the relevant features

Column	Type	Label	Description
STATION	text	Station	Unique station identifier (database key). For most stations, this is the state well number.
MSMT_DATE	timestamp	Water level measurement date (PST)	Date/time (in PST) the groundwater level measurement was collected
WLM_RPE	numeric	RPE for a specific water level measurement record	Reference point elevation used to collect the groundwater level measurement
WLM_GSE	numeric	GSE for a specific water level measurement record	Ground surface elevation at the well site
RPE_WSE	numeric	RPE to WSE	Depth to the water surface in feet below the reference point
GSE_WSE	numeric	GSE to WSE	Depth below the ground surface or the distance from the ground surface to the water surface in feet
WSE (target feature)	numeric	WS elevation	Water surface elevation in feet above mean sea level (NAVD88)

select the best ANN model, and the second one is applied to the resulting model from the hyperparameter optimizer for weight optimization.

3.1. Data preprocessing: principal component analysis (variables analysis)

In the first step of this study, we adopted the PCA as a primary step in the preprocessing of the data to minimize the dimension of the set of variables employed in the linear regression and, therefore, decrease the trained parameters of the neural networks.

Variable normalization (data reduction)

The data reduction in the PCA method aims to reduce heterogeneous variables, and it is required whenever the variables have several units of measurement and we want to assign equal importance to all features or whenever the dimension of the set of variables is very high. As in our case, we have heterogeneous variables, and we have proceeded to reduce them. There are several techniques applied to variable normalization, and the most commonly used one is to divide the values by the standard deviation implied in computing the Euclidean distance between individuals (5). Hence, the variables are reduced or centered. This computation allows more equitable role assignment (importance) to the whole variables (Duby & Robin 2006):

$$d^2(U_i, U_{i'}) = \sqrt{\sum_{j=1}^{p} \frac{(x_{ij} - \overline{x_{i'j}})^2}{\sigma_j^2}} \tag{5}$$

We computed a PCA of three components after the data normalization, and then we obtained the projection of the principal components' inertia percentages in Figure 6, which shows that the first and second components explain over 83% of the variance and more. Thus, the choice of three principal components was enough to explain most of the variability in the data. The projection of the contributions of the variables to the principal components illustrates the correlation matrix of features and principal components that allows selecting the features that correlate strongly with the most informative principal components, explaining most of the variability in the data. It appears clearly, according to the analysis of this graphic, that the most relevant variables are those projected with great dependence on the first and second components and present in Table 4. In this case, the lightest and darkest colors represent the most relevant parameters, guaranteeing a better distribution of the eigenvalues. Hence, the variables that follow the same direction as the groundwater level (WSE) are those that have a positive correlation and are represented by the lightest colors in the first component. Consequently, the training variables are reduced from 12 to 7 while keeping the date and the station parameters required for the time series analysis.

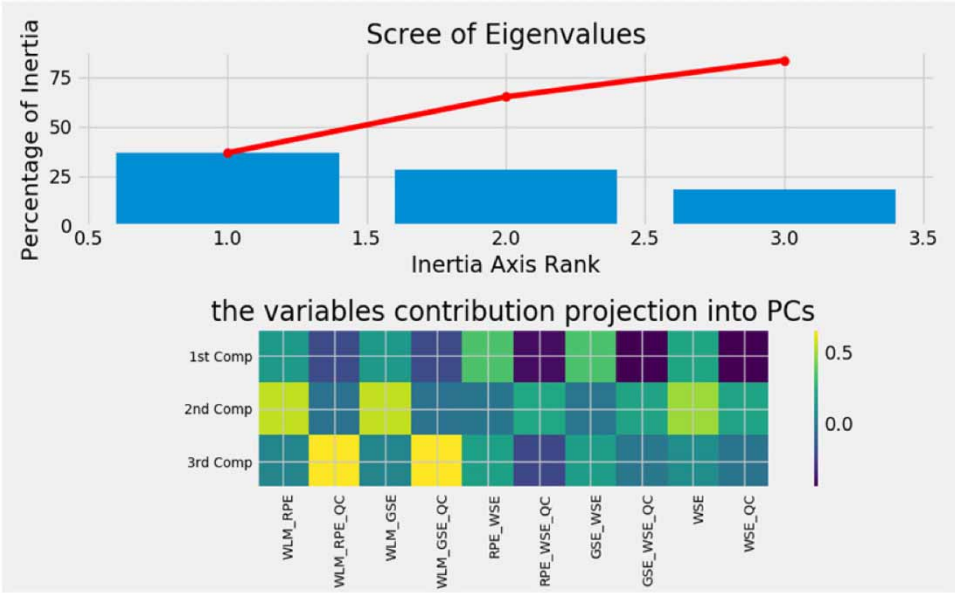


Figure 6 | PCA projection.

3.2. Implementation

In this subdivision, we expose the retrieved outcomes after training and testing both methods in the data selected from the monitoring well station '12N02E21Q003M'. Then we calculated various error measures for the efficiency comparison. After picking the most important features, we made forecasts by employing various ANNs, basic and combined, with the GAs for hyperparameters and weights optimization (basic ANN-MLP, ANN-XGA(HPO), ANN-XGA-HPO(WO) with the configuration above, based on the same training features, training period (5229 rows), and test period (about nine months from January 2021 to September 2021:271 days/ rows). Then we predicted the water surface elevation (WSE). For the evaluation, we have used the MSE as the loss function, the RMSE, the MAE, and the R_2-accuracy as metrics.

In the predictive basic ANN-MLP that we used in Figure 7, we implemented a sequential multi-layer perceptron model with a total number of neurons equal to 100. This model includes the rmsprop optimizer, an input layer of the relevant features, including a hidden layer with the activation function relu and 70 hidden units, a second hidden layer with the activation function relu and 29 hidden units, and an output layer (1 unit) with a linear activation function.

After training the extended GA mutation for the hyperparameter optimization of the ANN model using the R_2-accuracy as a fitness function, the set of a random selection of hyperparameters present in Table 5 for the population initialization, and the configuration present in Table 6, over ten generations and five solutions per population, we have obtained the

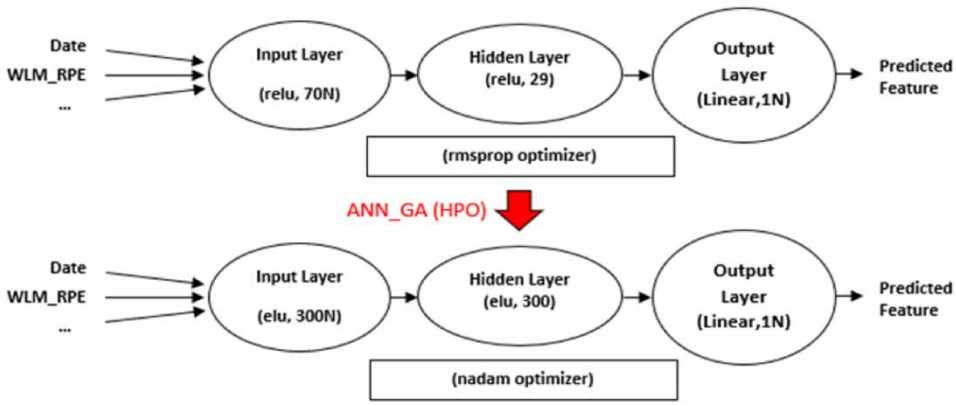


Figure 7 | Artificial neural networks' hyperparameters optimization.

Table 5 | Randomized hyperparameter of the population initialization and mutation

Column	Type	Description	Set of possible values
nb_neurons	numeric	Number of neurons of hidden layers	[20, 50, 100, 200, 300, 700]
nb_layers	numeric	Number of hidden layers	[1, 2, 3, 4]
activation	text	Activation function (hidden)	['relu', 'elu', 'tanh', 'sigmoid']
optimizer	text	Compilated optimizer	['rmsprop', 'adam', 'sgd', 'adagrad', 'adadelta', 'adamax', 'nadam']

Table 6 | Extended genetic algorithm (XGA-HPO)' parameters

Column	Type	Description	Value
nb_generations	numeric	Number of generations	10
nb_sol_per_pop	numeric	Number of solutions per population	5
Mutation_type	text	Mutation type	random
Crossover_type	text	Type of crossover	random
Fitness_function	numeric	Fitness function	R_2-accuracy
Solutions_retain	numeric	Percentage of retaining of solutions after each generation	40
mutation_percent	numeric	Mutation percentage of the genes	20

optimal solution (Optimized Artificial Neural Network: R_2-accuracy up to 99.92%) with the configuration illustrated in Figure 7:

After optimizing the hyperparameters for the initial feedforward neural network (ANN-XGA-HPO), we have performed an optimization of the weights using the second extended GA on the obtained solution using the configuration present in Table 7.

Training the GA with ANNs using large data samples increases the complexity of the processing, leading to exploding computational time even with reduced dimension. Thus, when we try to train the ANN with all the station's records present in the data set, it takes a very long processing time with an accuracy that is lower than training with specific station records, and it seems that the processing will not finish, especially with an increased number of generations. For these reasons, we have chosen to perform the training using small data sets and specific well station records. Figure 8 illustrates the different results generated by each model. The trends show that the ANN-XGA-HPO model (99.8%) with optimized hyperparameters is more precise than the basic ANN-MLP model (96.9%). Also, it is clearly shown that the predicted water surface elevation (WSE) in the studied well site using the ANN-HPO-WO model with optimized hyperparameters and weights is the most accurate and similar to the observed measurements, with an accuracy equal to (99.9%). Based on the curves, it seems that the amount of groundwater decreases significantly (discharge) with some increases (recharges) in the studied monitoring well in the period between May and September of 2021, which means that there is a critical need for water use optimization in the present and near future.

Table 7 | Extended genetic algorithm (XGA-WO)' parameters

Column	Type	Description	Value
nb_generations	numeric	Number of generations	10
nb_sol_per_pop	numeric	Number of solutions per population	5
Mutation_type	text	Mutation type	random
Crossover_type	text	Type of crossover	single point
Fitness_function	numeric	Fitness function	1/MSE
mutation_percent	numeric	Mutation percentage of the genes	10

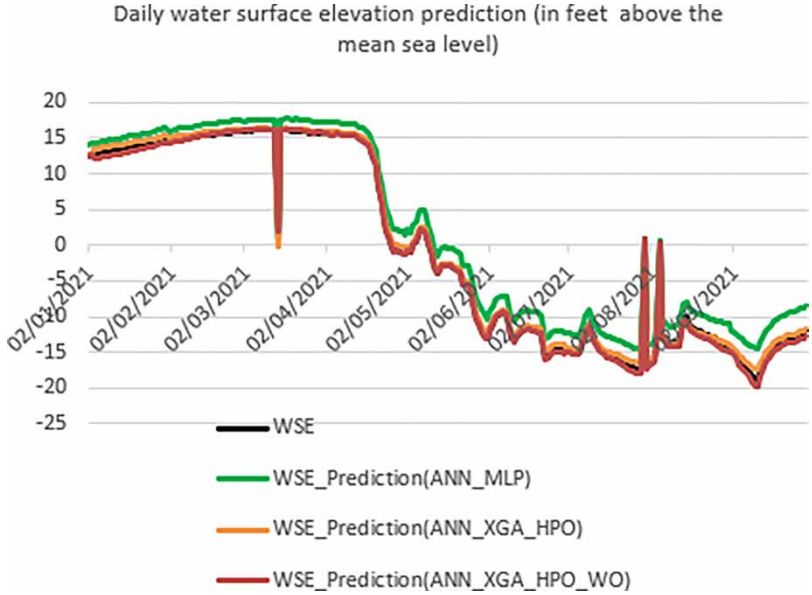


Figure 8 | Prediction of the water surface elevation (in feet above the mean sea level) in the monitoring well station '12N02E21Q003M' located in California.

Table 8 | Performance metrics output

	WSE		
Model	**RMSE**	**MAE**	**R_2 accuracy**
ANN-MLP	2.343	2.239	0.969
ANN_XGA_HPO	0.475	0.409	0.998
ANN_XGA_HPO_WO	0.423	0.325	0.999

In the comparison between predicted and observed values, the RMSE of the water surface elevation is lower for the ANN-XGA-HPO-WO model with both hyperparameters and weights optimization (0.423) than for the ANN-XGA-HPO model with only hyperparameters optimization (0.475), which is even lower than for the basic ANN-MLP model (2.343). The RMSE expresses the standard deviation of the residuals (errors) accumulated between observed and predicted values. In this case, both hyperparameters and weight optimization reduced the RMSE, and MAE errors, and increased the R_2 accuracy. By analyzing the performance evaluation outputs shown in Table 8, it appears that after both the hyperparameters and weights optimization using the extended GA, the performance of the ANN model had been empowered and outperformed the basic ANN-MLP model. Hence, the optimized model is precise, with an accuracy equal to (99.9%), and is reliable for modeling groundwater fluctuations. Consequently, ANN models are very suitable for computing complex groundwater systems, because of their ability to detect complex and nonlinear relationships such as soil water dynamics, especially when we empower them with data reduction tools like the PCA and optimization techniques like the GA.

Conclusively, monitoring wells and predicting the groundwater level constitute an essential step for controlling the water status and use in agriculture and industries. Modern methods measure the groundwater level in the nap in various aquifers. Thus, new information technologies make it possible to forecast the water status and the environment to adjust the groundwater use plan to climate change. Spending on this material, given that it is strongly designed and with good interpretation of the curves, enables us to purpose and better plan water use in short and over a long period of time.

4. CONCLUSION

Groundwater resources are exploited mostly by farmers to irrigate their plantations. Given the aggressive water demand and the drought related to climate change, a huge amount of groundwater is wasted because of evapotranspiration, which will

drive resources to decrease rapidly. Moreover, smart groundwater monitoring and supervision policies represent a primordial issue to overcome critical exploitations of underground water obtained from aquifers in the agricultural regions. Thus, smart tools for measuring and predicting groundwater level and depth can also be helpful for irrigation planning and water use monitoring. Moreover, from the obtained results, we can deduce that by integrating optimization and reduction techniques like PCA and GA in the data selection, hyperparameter initialization, and weight boosting, we can enhance the performance of the ANN model.

This paper introduces infrastructure for groundwater monitoring support through forecasting groundwater level changes, using an optimized ANN. The diversification of the data preprocessing using PCA, forecasting, and optimization techniques like the GA combined with ANNs would allow extracting a more suitable mechanism with high precision, for such precision-sensitive hydrological context towards a smart groundwater management system.

In perspective, this work could be improved by the implementation of a mobile system that can help in remotely monitoring the groundwater, throughout the changes in the hydrological cycle and climatic changes. Moreover, the proposed predictive approach could be extended using other optimization techniques and deployed in large data processing environments like Hadoop and Spark. Thus, integrating smartphones, weather remote sensing power, Bigdata execution environment, to the suggested framework describes the future expanse of automation. Therefore, the self-regulation of groundwater management processes involving ground and environmental parameter sensing devices derived from the Internet of Things and artificial intelligence would support water use-related decisions and facilitate groundwater level extraction and monitoring.

CONFLICT OF INTEREST

The authors declare no competing interests.

FUNDING

This research was funded by the National Center for Scientific and Technical Research of Morocco.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (https://data.cnra.ca.gov/dataset/continuous-groundwater-level-measurements).

REFERENCES

Alrashed, A. A. A. A., Gharibdousti, M. S., Goodarzi, M., de Oliveira, L. R., Safaei, M. R. & Bandarra Filho, E. P. 2018a Effects on thermophysical properties of carbon based nanofluids: experimental data, modelling using regression, ANFIS and ANN. *International Journal of Heat and Mass Transfer* **125**, 920–932.

Alrashed, A. A. A. A., Karimipour, A., Bagherzadeh, S. A., Safaei, M. R. & Afrand, M. 2018b Electro- and thermophysical properties of water-based nanofluids containing copper ferrite nanoparticles coated with silica: experimental data, modeling through enhanced ANN and curve fitting. *International Journal of Heat and Mass Transfer* **127**, 406–415.

Bagherzadeh, S. A., D'Orazio, A., Karimipour, A., Goodarzi, M. & Bach, Q. V. 2019 A novel sensitivity analysis model of EANN for F-MWCNTs–Fe_3O_4/EG nanofluid thermal conductivity: outputs predicted analytically instead of numerically to more accuracy and less costs. *Physica A: Statistical Mechanics and its Applications* **521**, 159–168.

Bahrami, M., Akbari, M., Bagherzadeh, S. A., Karimipour, A., Afrand, M. & Goodarzi, M. 2019 Develop 24 dissimilar ANNs by suitable architectures & training algorithms via sensitivity analysis to better statistical presentation: measure MSEs between targets & ANN for Fe–CuO/Eg–Water nanofluid. *Physica A: Statistical Mechanics and its Applications* **519**. https://data.cnra.ca.gov/dataset/continuous-groundwater-level-measurements/resource/84e02633-00ca-47e8-97ec-c0093313ddcd.

CDWR 2019 *Continious Groundwater Level Measurements*. California Department of Water Resources. https://data.cnra.ca.gov/dataset/periodic-groundwater-level-measurements%0Ahttps://data.ca.gov/dataset/continuous-groundwater-level-measurements.

Duby, C. & Robin, S. 2006 *Analyse en Composantes Principales*. Institut National Agronomique, Paris-Grignon, p. 80.

Edwards, E. C. & Guilfoos, T. 2021 The economics of groundwater governance institutions across the globe. *Applied Economic Perspectives and Policy* **43** (4), 1571–1594.

Eiben, A. E. & Smith, J. E. 2003 *Introduction to Evolutionary Computing Genetic Algorithms*. Natural Computing Series 45. Springer, Berlin, Heidelberg.

Gad, A. F. 2018 *Practical Computer Vision Applications Using Deep Learning with CNNs*. Apress, Berkeley, CA.

Ghasemi, A., Hassani, M., Goodarzi, M., Afrand, M. & Manafi, S. 2019 Appraising influence of COOH-MWCNTs on thermal conductivity of antifreeze using curve fitting and neural network. *Physica A: Statistical Mechanics and its Applications* **514**, 36–45.

Giwa, S. O., Sharifpur, M., Goodarzi, M., Alsulami, H. & Meyer, J. P. 2021 Influence of base fluid, temperature, and concentration on the thermophysical properties of hybrid nanofluids of alumina–ferrofluid: experimental data, modeling through enhanced ANN, ANFIS, and curve fitting. *Journal of Thermal Analysis and Calorimetry* **143** (6), 4149–4167.

Guerrien, M. 2003 L'intérêt de l'analyse en composantes principales (ACP) pour la recherche en sciences sociales. *Cahiers des Amériques Latines* **43**, 181–192.

Gupta, N. 2013 Artificial neural network. *Network and Complex Systems* **3** (1), 24–28.

Guzy, A. & Malinowska, A. A. 2020 State of the art and recent advancements in the modelling of land subsidence induced by groundwater withdrawal. *Water (Switzerland)* **12** (7), 2051.

Jolliffe, I. T. 2002 *Principal Component Analysis*, Vol. 19862. Springer-Verlag, New York.

Karimipour, A., Bagherzadeh, S. A., Goodarzi, M., Alnaqi, A. A., Bahiraei, M., Safaei, M. R. & Shadloo, M. S. 2018 Synthesized $CuFe_2O_4$/SiO_2 nanocomposites added to water/EG: evaluation of the thermophysical properties beside sensitivity analysis & EANN. *International Journal of Heat and Mass Transfer* **127**, 1169–1179.

Karimipour, A., Bagherzadeh, S. A., Taghipour, A., Abdollahi, A. & Safaei, M. R. 2019 A novel nonlinear regression model of SVR as a substitute for ANN to predict conductivity of MWCNT-CuO/water hybrid nanofluid based on empirical data. *Physica A: Statistical Mechanics and its Applications* **521**, 89–97.

Khosravi, R., Rabiei, S., Khaki, M., Safaei, M. R. & Goodarzi, M. 2021 Entropy generation of graphene–platinum hybrid nanofluid flow through a wavy cylindrical microchannel solar receiver by using neural networks. *Journal of Thermal Analysis and Calorimetry* **145** (4), 1949–1967.

Lall, U., Josset, L. & Russo, T. 2020 Annual review of environment and resources A snapshot of the world's groundwater challenges. *Annual Review of Environment and Resources* **45**, 171–194.

Moradikazerouni, A., Hajizadeh, A., Safaei, M. R., Afrand, M., Yarmand, H. & Zulkifli, N. W. B. M. 2019 Assessment of thermal conductivity enhancement of nano-antifreeze containing single-walled carbon nanotubes: optimal artificial neural network and curve-fitting. *Physica A: Statistical Mechanics and its Applications* **521**, 138–145.

Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B. & Agarwal, D. 2021 Surrogate optimization of deep neural networks for groundwater predictions. *Journal of Global Optimization* **81** (1), 203–231.

Panagopoulos, A. 2021a Energetic, economic and environmental assessment of zero liquid discharge (ZLD) brackish water and seawater desalination systems. *Energy Conversion and Management* **235**, 113957.

Panagopoulos, A. 2021b Study and evaluation of the characteristics of saline wastewater (brine) produced by desalination and industrial plants. *Environmental Science and Pollution Research* **29** (16), 23736–23749.

Panagopoulos, A. 2021c Techno-economic assessment of minimal liquid discharge (MLD) treatment systems for saline wastewater (brine) management and treatment. *Process Safety and Environmental Protection* **146**, 113957.

Peng, Y., Parsian, A., Khodadadi, H., Akbari, M., Ghani, K., Goodarzi, M. & Bach, Q. V. 2020 Develop optimal network topology of artificial neural network (AONN) to predict the hybrid nanofluids thermal conductivity according to the empirical data of Al_2O_3 – Cu nanoparticles dispersed in ethylene glycol. *Physica A: Statistical Mechanics and its Applications* **549**, 124015.

Ringnér, M. 2008 What is principal component analysis? *Nature Biotechnology* **26** (3), 303–304.

Rowe, R. C. & Colbourn, E. A. 2003 Neural computing in product formulation. *The Chemical Educator* **8** (3), 1–81.

Safaei, M. R., Hajizadeh, A., Afrand, M., Qi, C., Yarmand, H. & Zulkifli, N. W. B. M. 2019 Evaluating the effect of temperature and concentration on the thermal conductivity of ZnO-TiO_2/EG hybrid nanofluid using artificial neural network and curve fitting on experimental data. *Physica A: Statistical Mechanics and its Applications* **519**, 209–216.

Sahoo, S., Russo, T. A., Elliott, J. & Foster, I. 2017 Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research* **53** (5), 3878–3895.

Schmidhuber, J. 2015 Deep learning in neural networks: an overview. *Neural Networks* **61**, 85–117.

Taylor, C. J. & Alley, W. M. 2001 *Ground-water-level Monitoring and the Importance of Long-Term Water-Level Data*, Vol. 1217. US Geological Survey Circular, pp. 1–68.

Wen, X., Feng, Q., Deo, R. C., Wu, M. & Si, J. 2017 Wavelet analysis-artificial neural network conjunction models for multi-scale monthly groundwater level predicting in an arid inland river basin, northwestern China. *Hydrology Research* **48** (6), 1710–1729.

Wunsch, A., Liesch, T. & Broda, S. 2021 Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences* **25** (3), 1671–1687.

First received 13 December 2021; accepted in revised form 31 March 2022. Available online 16 April 2022