Prediction of groundwater flow in shallow aquifers using artificial neural networks in the northern basins of Algeria

N. Guezgouz, D. Boutoutaou and A. Hani

ABSTRACT

Prediction of groundwater flow fluctuations is considered an important step in understanding groundwater systems at this scale and facilitating sustainable groundwater management. The objective of this study is to determine the factors that influence and control groundwater flow fluctuations in a specific geomorphologic situation, by developing a forecasting model and examining its potential for predicting groundwater flow using limited data. Models for prediction of groundwater flow are developed based on artificial neural networks (ANNs). Neural networks with different numbers of hidden layer neurons were developed using climatic and geomorphological characteristics as input variables, giving predicted groundwater flow as the output. To evaluate enhanced performance models, several regression statistical parameters are compared. As an example, relative mean square error in groundwater flow prediction by ANN and correlation coefficient are 0.015 and 97%, respectively. The results of the study clearly show that ANNs can be used to predict groundwater flow in shallow aquifers of northern Algeria with reasonable accuracy even in the case of limited data.

Key words | groundwater, limited data, model, northern Algeria, prediction

N. Guezgouz (corresponding author)
Department of Biology, Faculty of Natural and Life Sciences,
University of Mohamed Cherif Messadia,
Souk-Ahras,
Algeria
E-mail: *n.guezgouz@univ-soukahras.dz*

D. Boutoutaou
Department of Civil Engineering and Hydraulics, Faculty of Applied Science,
Kasdi Merbah Ouargla University,
30000 Oaurgla,
Algeria

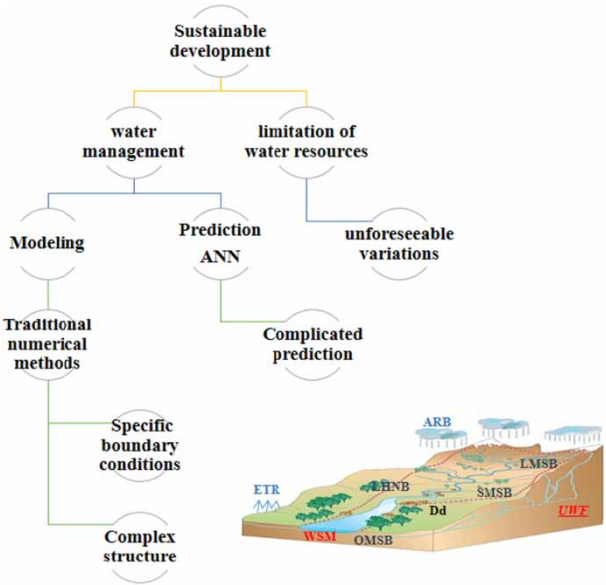
A. Hani
Water Resources and Sustainable Development Laboratory,
University of Badji Mokhtar,
Annaba,
Algeria

HIGHLIGHTS

- Combine hydrological and climatic data to estimate groundwater flows.
- Test the performance of ANN's models to understand the behavior of groundwater.
- Large-scale groundwater flow modeling for better management of water resources.
- Proposal of a predictive model for a global vision of the distribution of groundwater.
- Determining the order of importance of indicators that can influence groundwater flows.

doi: 10.2166/wcc.2020.067

GRAPHICAL ABSTRACT



INTRODUCTION

Integrated water resources management is a systematic process for sustainable development, allocation and monitoring of water resources viewed as both a geomorphological influence and a climatic variation. This conceptual model interprets the two systems through three components including the watershed nature, the stream characteristics and the rainfall, influencing groundwater flows.

To assist water planners and managers to gain adequate knowledge and understanding of the relationships between the response variables and water resources mobilization, there is a need to use a proper methodology to define the effective response variable influencing the attractiveness of water resources mobilization.

In recent years, the artificial neural networks (ANNs) models have been successfully applied to hydrological processes, such as rainfall–runoff modeling (Minns & Hall 1996) and rainfall forecasting (Lallahem & Mania 2003) and in water resources context, the ANN has been used for water quality parameters (Maier & Dandy 2000), forecasting of water demand (Liu *et al.* 2003), stream flow forecasting (Change *et al.* 2003), prediction of rainfall–runoff relationship (Change *et al.* 2003; Riad *et al.* 2004), coastal aquifer management (Albaradeyi *et al.* 2011), stream flow modeling (Coulibaly *et al.* 2000) and reservoir operation problems (Hornik *et al.* 1989). Hornik *et al.* (1989) showed how ANN could be applied to different problems in civil engineering, while Maier & Dandy (2000) reviewed several papers dealing with the use of neural network models for the prediction and forecasting of water resources variables.

An ANN can effectively establish the relationship between the input and output variables without considering the detailed physical process, which attracts increasing attention in terms of predicting the groundwater flow.

A back-propagation feed-forward multilayer perceptron (MLP) with sigmoidal-type transfer functions is the most popular neural network architecture due to its high performance compared to the other ANNs (Lippmann 1987).

This study aims to establish a modeling relationship between groundwater flow and response variables in shallow aquifers in the north of Algeria, characterizing their priorities, to better manage water resources, especially under climate change and the first rains delay, causing ugly impacts on agricultural activity which uses the rainfall for irrigation.

MATERIALS AND METHODS

Study area

The study basins are situated in the north of Algeria (Figure 1). They are bordered by Morocco from the west, the Algerian Sahara basin from the south, Tunisia from the east and the Mediterranean Sea from the north. The total area of the northern Algeria river basins is about 480,000 km^2, comprising 17 river basins and 224 sub-basins.

Water resources in the study area are vulnerable to the fast-growing demand of urban and rural populations, demand of economic sectors including agriculture, industry and public institutions. Groundwater from shallow aquifers in northern Algeria is used primarily to irrigate vegetable crops, over an area exceeding 1 million hectares. It is also an important source of drinking water in rural areas through traditional wells.

Data description

Groundwater flow data and response variables were implemented in the ANN model using the software package of STATISTICA 8 (Serial: STA862D175437Q). The Arc-Hydro Toolbox was used to extract geomorphometric land surface variables and features from Digital Elevation Models (DEMs). It comprises a series of Python/NumPy processing functions, presented through an easy-to-use graphical menu in the widely used ArcGIS package. Climatic data were sourced from government agencies as independent datasets (each case is independent) for the observed sub-basins. The response variables were:

- groundwater velocity (GWV) (mm yr^{-1});
- area of watershed (AWS) (km^2);
- drainage density (Dd) (km km^{-2});
- order of the main stream in the basin (OMSB) (–);
- length of the main stream in the basin (LMSB) (km);
- slope of the main stream in the basin (SMSB) (m km^{-1});
- length of hydrographic network in the basin (LHNB) (km);
- hydro-morphological coefficient of the basin (HMCB) (–);
- annual rainfall in the basin (ARB) (mm yr^{-1});
- stream water flow (SWF) (mm yr^{-1}); and
- evapotranspiration in the basin (ETR) (mm yr^{-1}).

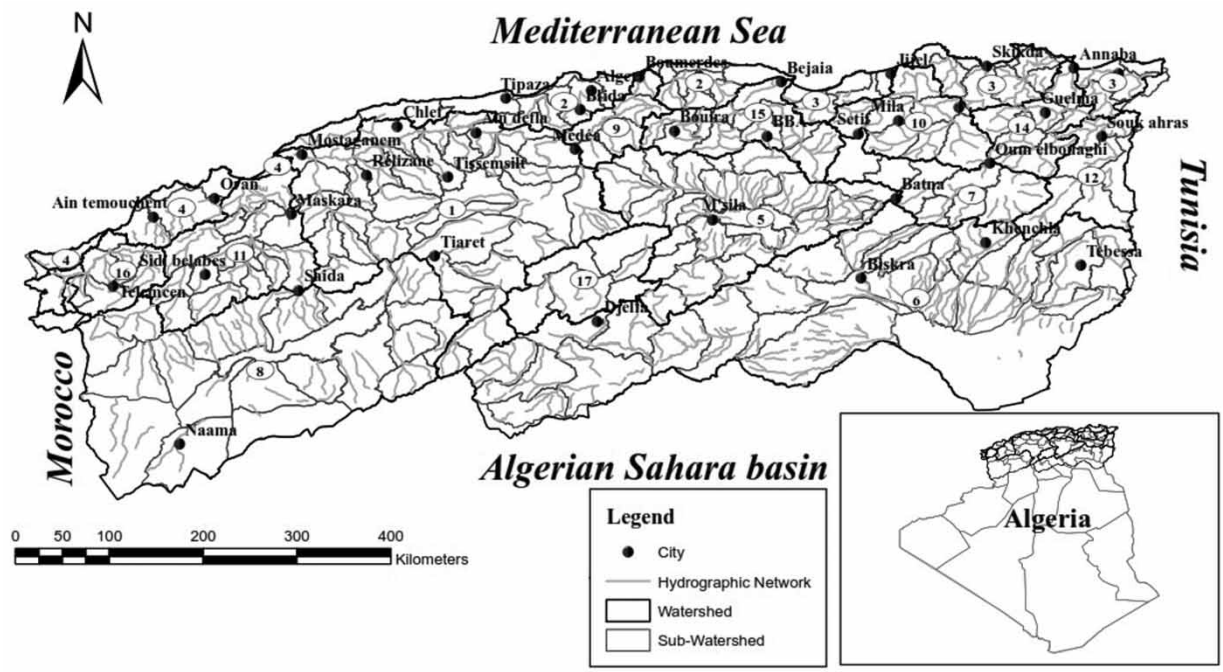


Figure 1 | Location of the study basins: northern Algeria.

The variables representing the response category are considered as the possible input variables while the target output variable is GWV. All input variables will be compared with expert opinion and judgment ranking to assess the performance of the conceptual model.

Evaluation criteria

A variety of verification criteria that could be used for the evaluation and intercomparison of different models was proposed by the World Meteorological Organization (WMO). They fall into two groups: graphical indicators and numerical performance indicators (W.M.O 1975). The root mean square error (RMSE) and the correlation coefficient (R^2) (Legates & Mc Cabe 1999) are chosen for the present study, given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - \widehat{Q_i})^2} \tag{1}$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Q_i - \widehat{Q_i})(\widehat{Q_i} - \overline{\widehat{Q_i}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \widehat{Q_i})^2}\sqrt{\sum_{i=1}^{n}(\widehat{Q_i} - \overline{\widehat{Q_i}})^2}}\right]^2 \tag{2}$$

where Q_i the observed groundwater velocity value; $\widehat{Q_i}$ is the predicted groundwater velocity value; $\bar{Q}_i$ is the mean value of Q_i values; $\overline{\widehat{Q_i}}$ is the mean value of $\widehat{Q_i}$ values and n is the total number of data values.

The RMSE gives a quantitative indication of the network error. It measures the deviation of the predicted values from the corresponding observed values of target output (Lallahem *et al.* 2004; Hani *et al.* 2006). The RMSE was used to compare the performance of MLP with other common types of ANNs, such as the Radial Basis Function (RBF).

The R^2 value is an indicator of how well the network fits the data and accounts for the variability with the variables specified in the network. A value of R^2 above 90% refers to a very satisfactory model performance. Values range between 80 and 90% indicates the unsatisfactory model (Lallahem & Mania 2003; Riad *et al.* 2004). The ideal value for RMSE is zero and for R^2 is unity.

Architecture of the network

ANN models are mathematical tools, capable of modeling extremely complex functions and a wide spectrum of challenging problems (Liu *et al.* 2003). They constitute a computational approach inspired by the human nervous system. The processing units of an artificial neural network are called neurons, which are arranged into layers. Neurons between layers are connected by links of variable weights. The number of neurons in a hidden layer is decided after training and testing. Training of ANN consists of showing example inputs and target outputs to the network and iteratively adjusting internal parameters based on performance measures. Multilayered networks, trained by back propagation (Rumelhart *et al.* 1986) are currently the most popular and efficient (Hagan *et al.* 1996). They have been used in this study.

The most popular neural network models are the RBF and the MLP. The MLP is a layered feed-forward network, which is typically trained with BFGS (Broyden Fletcher Goldfarb Shanno) quasi-Newton back propagation (Broyden 1970; Shanno David 1970) and SCG (Scaled Conjugate Gradient) back propagation. The MLP is simple, robust and very powerful in pattern recognition, classification and mapping. MLP is capable of approximating any measurable function from one finite-dimensional space to another within a desired degree of accuracy (Hornik *et al.* 1989).

In this work, a feed-forward MLP network with a back-propagation algorithm was chosen to model the system.

The network processes are an input vector consisting of possible variables, including *AWS*, *Dd*, *OMSB*, *LMSB*, *SMSB*, *LHNB*, *HMCB*, *ARB*, *SWF* and *ETR*. This input vector generates an output vector which is *GWV*. The MLP network can be represented by the following compact form:

$$\{GWV\} = \mathrm{ANN}\,[AWS,\ Dd,\ OMSB,\ LMSB,\ SMSB,\ LHNB,\ HMCB,\ ARB,\ SWF,\ ETR]$$

A schematic diagram of the neural network is shown in Figure 2. It shows a typical feed-forward structure with signals flowing from input nodes, forward through hidden nodes and eventually reaching the output node. The input

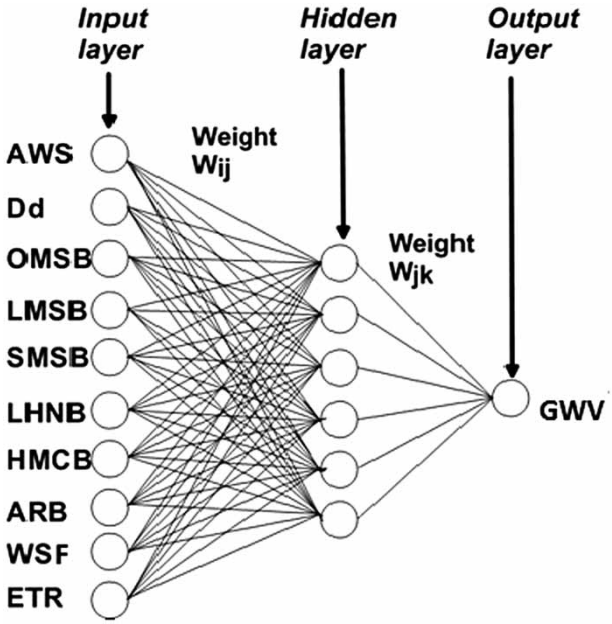


Figure 2 | Architecture of the neural network model in this study.

layer is not really neural at all; these nodes simply serve to introduce the standardized values of the input variables to the neighboring hidden layer without any transformation. The hidden and output layer nodes are each connected to all of the nodes in the preceding layer. However, the nodes in each layer are not connected to one another. A numeric weight is associated with each of the inter-node connections. A weight of W_{ij} represents the strength of connections of nodes between the input and hidden layer, while W_{jk} represents the strength of connections of nodes between the hidden and output layers.

Each hidden node (j) receives signals from every input node (i), comprising standardized values $\overline{(X_i)}$ of an input variable, where various input variables from (Xmin) to (Xmax) have different measurement units and span different ranges. $\overline{X_i}$ is expressed as follows:

$$\overline{X_i} = \frac{X_i - X\min(i)}{X\max(i) - X\min(i)} \tag{3}$$

Each signal comes via a connection that has a weight (W_{ij}). The net integral of incoming signals to a receiving hidden node (NET_j) is the weighted sum of the input signals, $\overline{X_i}$, and the corresponding weights, W_{ij}, plus a constant reflecting the node threshold value (TH_j):

$$\mathrm{NET}_j = \sum_{i=1}^{n} X_i W_{ij} + \mathrm{TH}_j \tag{4}$$

The net incoming signals to a hidden node (NET_j) are transformed to an input (O_j) from the hidden node by using a non-linear transfer function (f) of the sigmoid type, given by the following equation form:

$$O_j = f(\mathrm{NET}_j) = \frac{1}{1 + e^{-\mathrm{NET}_j}} \tag{5}$$

O_j passes as a signal to the output node (k).

The net entering signals to an output node (NET_k) are given by

$$\mathrm{NET}_k = \sum_{i=1}^{n} O_j W_{jk} + \mathrm{TH}_k \tag{6}$$

The net incoming signals of an output node (NET_k) are transformed using the sigmoid type function to a standardized or scaled output ($\bar{O}_k$), that is:

$$\bar{O}_k = f(\mathrm{NET}_k) = \frac{1}{1 + e^{-\mathrm{NET}_k}} \tag{7}$$

Then, $\bar{O}_k$ is standardized to produce the target output:

$$O_k = \bar{O}_k\,(O\max_k - O\min_k) + O\min_k \tag{8}$$

Riad *et al.* (2004) explained that the sigmoid function should be continuous, differentiable and bounded from above and below in the range [0,1]. The calculated error between the observed actual value and the predicted value of the dependent variable is back propagated through the network and the weights are adjusted. The cyclic process of feed forward and error back propagation is repeated until the verification error is minimal (Liu *et al.* 2003).

Calibration and verification of the model

In the case that limited datasets are available, cross-verification can be used as a stopping criterion to determine the optimal number of hidden layer nodes (Braddock *et al.* 1997) while avoiding the risk of over training. Cross-verification is a technique commonly used in ANN models and has a significant impact on the division of data (Burden *et al.* 1997). It aims to train the network using one set of data and to check performance against a verification set not used in training. This examines the ability of the network to generalize properly by observing whether the verification error is reasonably low. The training will be stopped when the verification error starts to increase (Figure 3; Lallahem & Mania 2003). The database was divided into training, cross-verification and testing. For the ANN models described in this paper, 50% of the available data were used for training, 25% were used for verification and 25% to test the validity of network prediction (Lallahem *et al.* 2004).

Setup of the model inputs

ANN models have the ability to determine which inputs are critical. They are useful mainly for complex problems where the number of potential inputs is large and where *a priori* knowledge is not available to determine appropriate inputs (Lachtermacher & Fuller 1994). In this study, a sensitivity analysis can be carried out to identify the importance of the input variables.

This indicates which variables are considered to be most useful to be retained by the ANN model. The ANN model removes the input variables with low sensitivity. The sensitivity is presented by the Ratio and Rank. The Ratio reports the relation between the Error and the Baseline Error (i.e. the error of the network if all variables are 'available'). The Rank simply lists the variables in the order of their importance.

RESULTS AND DISCUSSION

In the northern Algeria basins, groundwater flow is driven by stream flow, annual precipitation in the basin, drainage density and other various geomorphological variables. Stream flow has produced a root of the limited available groundwater flow and is assured by precipitation.

The types of considered networks are MLP with two back-propagation algorithms (BFGS and SCG) and RBF. During the analysis, many other networks were tested. The best optimal ANN model found is MLP (BFGS 137) with four hidden nodes and a smaller error (0.015) than the other types of ANN networks tested (Table 1).

Verification of the model demonstrates a good fit to the available data, RMSE values for training, verification and testing are consistently small in magnitude, indicating that

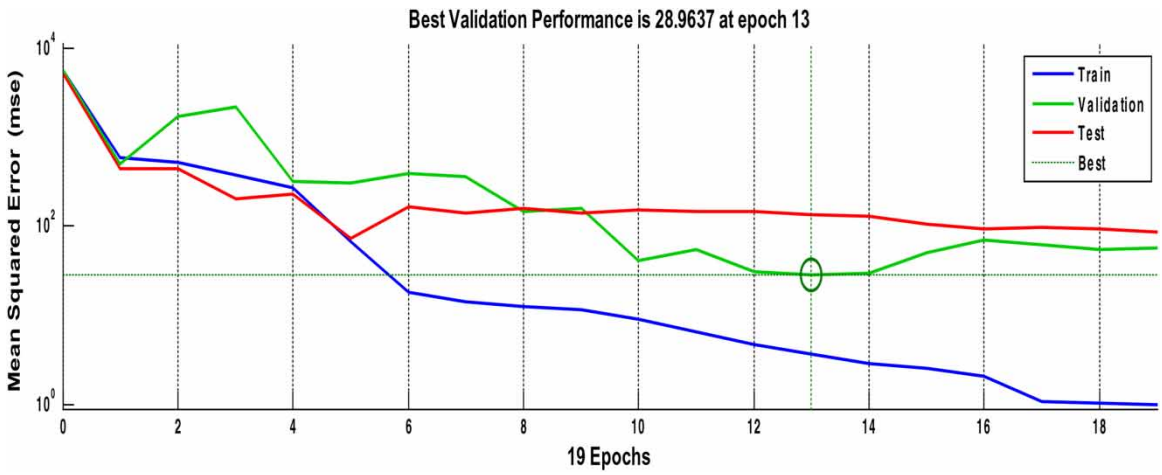


Figure 3 | Validation performance of training, validation and testing values.

Table 1 | RMSE for different ANNs

ANN	Architecture	RMSE
RBF	10-8-1	0.034
$MLP_{(BFGS\ 103)}$	10-6-1	0.029
$MLP\ (_{BFGS\ 137})$	10-4-1	0.015

the data subsets are from the same population (Jalala *et al.* 2011; Table 2). In addition, the correlation coefficient for each phase exceeds 97% which shows a close agreement between the observed and predicted groundwater velocity (Figure 4).

Table 2 | Regression statistical parameters for the target output (UWF)

ANN	Training	Validation	Testing
Data mean	18.0650	26.3087	44.3786
Standard deviation	16.1953	21.9924	52.8131
RMSE	0.000107	0.014525	0.014910
Correlation	0.997895	0.970188	0.970215

Samples: Train, Test.

The model training error for the independent cases is shown in Figure 5. It shows the RMSE on the training, testing and verification subsets of the independent cases at the end of the last iterative training run. The graph indicates that the range of RMSE of independent cases for both training, testing and verification is very small (Al-Mahallawi *et al.* 2011; Jalala *et al.* 2011). The ANN sensitivity analysis of response variables in both training and verification phases (Table 3) indicates that stream flow is the most important variable followed by the efficiency of annual precipitation.

The policy interventions according to their order in the verification phase are stream flow, annual precipitation, drainage density, hydro-morphological coefficient, slope of the main stream, evapotranspiration, area of watershed, length of hydrographic network, order of the main stream and length of the main stream in the basin. The results of the ANN model and expert opinion (Table 4) are similar only in ranking the first (stream flow), the second (annual precipitation), the third (drainage density) and the fifth intervention (slope of the main stream) while they differ in ranking the remaining variables.

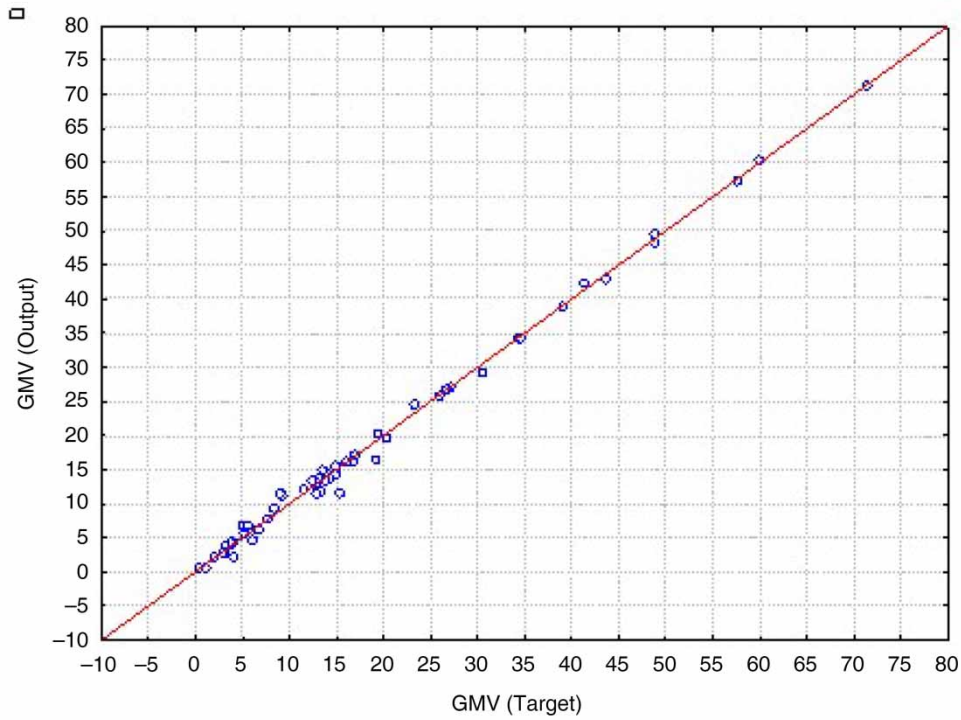


Figure 4 | Predicted GWV versus observed GWV (mm yr^{-1}).

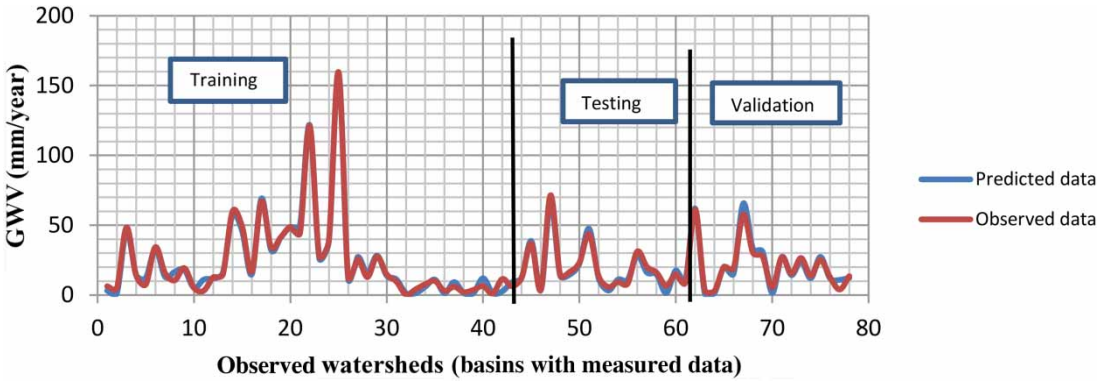


Figure 5 | Variation between observed and predicted groundwater velocities.

Table 3 | Sensitivity analysis of independent input variables in MLP (BFGS 137 and BFGS 103, respectively)

		AWS	LHNB	Dd	SMSB	LMSB	OMSB	ARB	SWF	ETR	HMCB
BFGS 137	Ratio	4.38	2.65	11.89	7.37	1.39	1.65	23.10	36.14	6.74	11.31
	Rank	7	8	3	5	10	9	2	1	6	4
BFGS 103	Ratio	1.73	1.28	5.38	3.16	3.55	1.19	6.20	18.43	3.96	2.86
	Rank	8	9	3	6	4	10	2	1	5	7

Table 4 | Ranking of input variables via expert opinion and judgment

	AWS	LHNB	Dd	SMSB	LMSB	OMSB	ARB	SWF	ETR	HMCB
Rank	6	4	3	5	10	9	2	1	7	8

CONCLUSION

In this study, the factors that influence and control groundwater velocity in a specific geomorphological and climatic situation were determined and used to develop ANN models for forecasting groundwater stocks for different watersheds in northern Algeria.

The obtained results indicate that an MLP network proved to be the best ANN structure to model and predict the relationship between response variables and groundwater velocity in the northern Algeria basins. The results are in good agreement with previous related studies done with datasets of longer duration. Therefore, it can be concluded that an ANN is an effective tool for forecasting groundwater velocities for the purposes of groundwater management, even though only limited data samples were available.

The model also supports the Integrated Water Resources Management (*IWRM*) approach by indicating that stream flow and inter-annual precipitation are the strongest controls on *GWV*.

Further investigations are needed to understand how alternative ANN architectures and training algorithms perform in data-poor situations. There is also considerable scope to implement other soft computing methods such us hydrogeological models to forecast groundwater velocities.

ACKNOWLEDGEMENTS

We would like to thank Mr Graham Malcom from Imperial Academia, UK, for providing language help.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Albaradeyi, I., Hani, A. & Shahrour, I. 2011 WEPP and ANN models for simulating soil loss and runoff in a semi-arid Mediterranean region. *Environmental Monitoring Assessment* **180**, 537–556.

Al-Mahallawi, K., Mania, J., Hani, A. & Shahrour, I. 2011 Using of neural networks for the prediction of nitrate groundwater contamination in rural and agricultural areas. *Environmental Earth Sciences* **65**, 917–928.

Braddock, R. D., Kremmer, M. L. & Sanzogni, L. 1997 Feed-forward artificial neural network model for forecasting rainfall run-off. In *Proceedings of the International Congress on Modeling and Simulation (Modsim)*, Hobart, Australia. The Modeling and Simulation Society of Australia Inc., pp. 1653–1658.

Broyden, C. G. 1970 The convergence of a class of double-rank minimization algorithms. *IMA Journal of Applied Mathematics* **6** (1), 76–90.

Burden, F. R., Brereton, R. G. & Walsh, P. T. 1997 Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. *Analyst* **122** (10), 1015–1022.

Change, L. C., Change, F. J. & Chiange, Y. M. 2003 A two-step-ahead recurrent neural network for stream-flow forecasting. *Hydrological Processes* **18**, 81–92. doi:10.1002/hyp.1313.

Coulibaly, P., Anctil, F. & Bobée, B. 2000 Daily reservoir in flow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology* **230**, 244–257.

Hagan, M. T., Demuth, H. B. & Beale, M. H. 1996 *Neural Network Design*, 1st edn. PWS Publishing Co., Boston, MA, USA. ISBN: 0-53494332-2.

Hani, A., Lallahem, S., Mania, J. & Djabri, L. 2006 On the use of finite-difference and neural network models to evaluate the impact of underground water overexploitation. *Hydrological Processes* **20**, 4381–4390.

Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feed forward networks are universal approximators. *Neural Networks* **2**, 359–366. doi:10.1016 /0893- 6080(89)90020-8.

Jalala, S., Hani, A. & Shahrour, I. 2011 Characterizing the socio-economic driving forces of groundwater abstraction with artificial neural networks and multivariate techniques. *Water Resource Management* **25**, 2147–2175.

Lachtermacher, G. & Fuller, J. D. 1994 Back propagation in hydrological time series forecasting. In: *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, Vol. 3 (K. W. Hipel, A. I. McLeod, U. S. Panu & V. P. Singh, eds.). Springer, Dordrecht, pp. 229–242.

Lallahem, S. & Mania, J. 2003 A nonlinear rainfall-runoff model using neural network technique: example in fractured porous media. *Mathematical and Computer Modeling* **37**, 1047–1061. doi:10.1016/S0895 7177(03) 00117-1.

Lallahem, S., Mania, J., Hani, A. & Najjar, Y. 2004 On the use of neural networks to evaluate groundwater levels in fractured media. *Journal of Hydrology* **307** (1–4), 738–744.

Legates, D. R. & Mc Cabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35** (1), 233–241.

Lippmann, R. P. 1987 An introduction to computing with neural nets. *ASSP Magazine, IEEE* **4** (2), 4–22. doi:10.1109/MASSP. 1987.1165576.

Liu, J., Savenije, H. H. G. & Xu, J. 2003 Forecast of water demand in Weinan City in China using WDF-ANN model. *Physics and Chemistry of the Earth* **28**, 219–224.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: are view of modeling issues and applications. *Environmental Modeling & Software* **15**, 101–124.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall-runoff models. *Hydrological Sciences* **41** (3), 399–417. doi:10.1080/02626669609491511.

Riad, S., Mania, J., Bouchaou, L. & Najjar, Y. 2004 Predicting catchment flow in a semi-arid region via an artificial neural network technique. *Hydrological Processes* **18**, 2387–2393. doi:10.1002/hyp.1469.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. 1986 Learning internal representations by error propagation. In: *Paralleled Distributed Processing. Explorations in the Microstructure of Cognition*, Vol. 1 (D. E. Rumelhart & J. L. McClelland, & the PDP Research Group, eds.). The MIT Press, Cambridge, MA, pp. 318–362.

Shanno David, F. 1970 Conditioning of quasi-Newton methods for function minimization. *Mathematical Computing* **24** (111), 647–656.

World Meteorological Organization 1975 *Inter-comparison of conceptual models used in operational hydrological forecasting*, W.M.O, technical series. *Water Resources Research* **27** (9), 2415–2450.

First received 17 February 2020; accepted in revised form 12 June 2020. Available online 14 July 2020