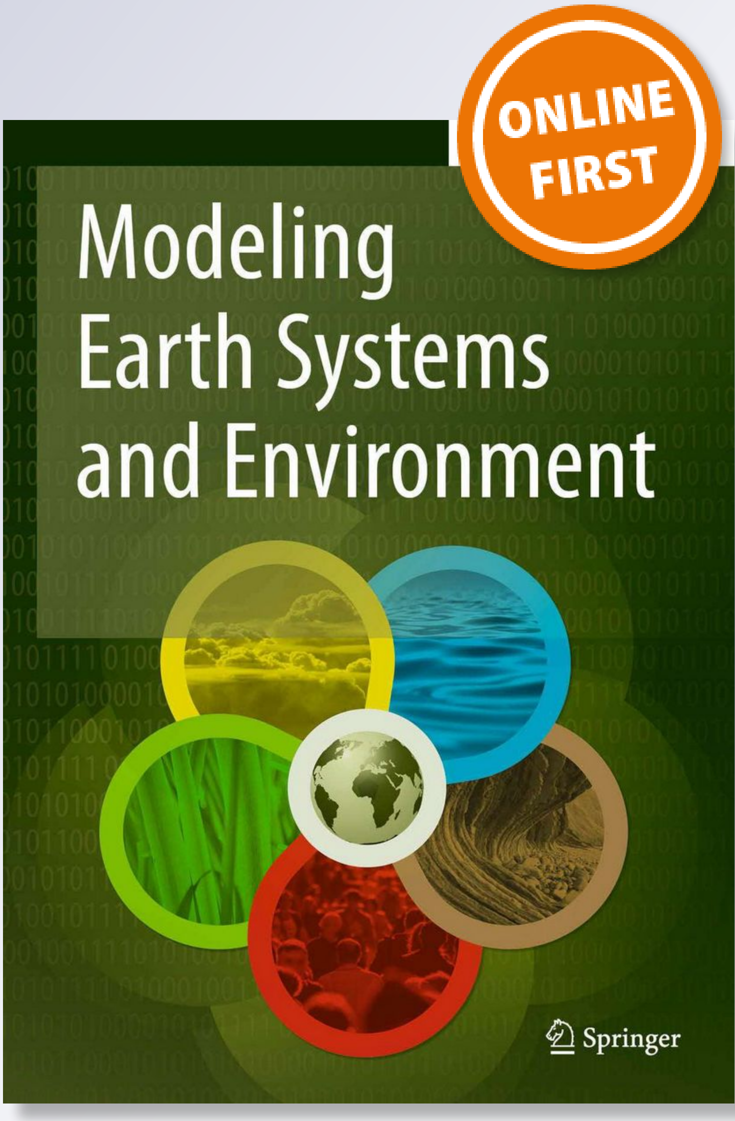
Application of automatic relevance determination model for groundwater quality index prediction by combining hydro-geochemical and geo-electrical data

Saumen Maiti, Anasuya Das, Rhythm Shah & Gautam Gupta

Modeling Earth Systems and Environment

ISSN 2363-6203

Model. Earth Syst. Environ.
DOI 10.1007/s40808-017-0369-x



Springer

Model. Earth Syst. Environ.
DOI 10.1007/s40808-017-0369-x

ORIGINAL ARTICLE

Application of automatic relevance determination model for groundwater quality index prediction by combining hydro-geochemical and geo-electrical data

Saumen Maiti[1] · Anasuya Das[1] · Rhythm Shah[1] · Gautam Gupta[2]

Received: 17 August 2017 / Accepted: 18 August 2017

Abstract In this study, an automatic relevance determination-based Bayesian neural network (ARD-BNN) approach is employed combing hydro-geochemical and geo-electrical information to predict groundwater quality index (GWQI) of coastal Maharashtra, India. In the first step, to incorporate geo-electrical information for GWQI forward computation, the weight corresponding to true resistivity of the earth layer is estimated via ARD-BNN modelling using training sample generated for groundwater classification. In the second step, incorporating earth resistivity weight information, a total of 1500 training samples are created honouring World Health Organisation (WHO) guidelines for GWQI prediction. Prior to actual data analysis, we explored the algorithm on GWQI variable series assorted with different level of complex (red) noise to examine the bounds of network hyper-parameter. What distinguishes our approach from previous approach for ARD-BNN optimization is that we seek to develop a mechanism which allows specific weight estimation and provides insight into which hyper parameters and their bounds are appropriate to predict GWQI from noise intervened GWQI data. The model shows excellent performance between predicted and computed GWQI both in training (trn) and test (tst) data with Pearson's correlation coefficient ($r_{trn} \sim 0.91$ and $r_{tst} \sim 0.90$), root-mean-squared-error (RMSE) error ($RMSE_{trn} \sim 1.2$ and $RMSE_{tst} \sim 1.4$), reduction of error (RE) ($RE_{trn} \sim 0.98$ and $RE_{tst} \sim 0.97$), and index of agreement (IA) ($IA_{trn} \sim 0.95$ and $IA_{tst} \sim 0.93$). Red noise analysis shows that the ARD-BNN model is robust up to the noise level of 20% in the input variable for GWQI prediction. The network adopted relevance analysis to indicate the relative importance of input parameter in the prediction of GWQI via ARD-BNN modelling which showed that chloride [Cl^-], [pH], [HCO_3^-] and sodium [Na^+] are dominant while considering various simulations for characterizing the GWQI of coastal Maharashtra. The approach used here, could be useful in understanding the relative contribution and/or modelling pollution source in many other environmental applications.

Keywords Bayesian neural networks · Automatic relevance determination · Groundwater quality parameter · Groundwater quality index · Konkan

Introduction

Groundwater is the most important natural resource used for drinking, irrigation and variety of other domestic purposes. The assessment of groundwater quality index (hereafter GWQI) at any place is an imperative task in hydrology. Prediction of GWQI is a difficult task due to uncertainty stemming from spatial variability, measurement noise, modelling error, statistical error, and non-linear/complex relationship among input groundwater quality variables with predicted variables (Brown 1998). Researchers used hydro-geochemical information for the assessment of groundwater (Todd 1959; Hem 1970; Fetter 1988; Schoeller 1959, 1967; Brown 1998; Brown et al. 1970; Pradhan et al. 2001; Singh et al. 2007; Stigter et al. 2006; Magesh et al. 2013 and many others). Horton (1965) introduced water quality index (WQI) by using indices so that a large number of water quality data

✉ Saumen Maiti
saumen_maiti2002@yahoo.co.in

[1] Department of Applied Geophysics, Indian Institute of Technology (Indian School of Mines), Dhanbad 826004, India

[2] Indian Institute of Geomagnetism (DST), Navi Mumbai 410218, India

Published online: 29 August 2017

can be summarised to a single numerical value. Using the concept many other researchers used WQI for groundwater quality studies (Ramakrishnaiah et al. 2009; Vasanthavigar et al. 2010; Krishna kumar et al. 2014).

Recently data driven modeling and/or neural network modelling based on a variety of algorithms has been found to be an alternative prospective scheme for water quality assessment (e.g. Vadiati et al. 2016; Gazzaz et al. 2012; Sadat-Noori et al. 2014; Wang and Lu 2006; Zou and Wang 2010; Sahu et al. 2011; Najah et al. 2013). Particularly, the widespread acceptance of neural network modeling in Bayesian framework is due to the fact that it prevents overfitting and provides additional means to measure the uncertainty in output prediction (Bishop 1995; Nabney 2004). The BNN technique has been used widely in various disciplines of applied science. For example, Lopez et al. (2005) used ARD-BNN for modeling solar energy and assessed the importance of input parameters for output prediction. Wang and Lu (2006) used ARD for ozone level modeling and assessed the relative influence of input climatic variables for output prediction. Calster et al. (2008) used ARD-BNN for tumor studies prior to surgery and have shown the way to assess the sensitivity of input parameters. Hippert and Taylor (2010) used ARD-BNN for short-term load forecasting. They have described the Bayesian neural network method and evidence approximation in compact form and shown that Bayesian method could be potential tool for solving problem related to electrical power. Maiti et al. (2012) used Hybrid Monte Carlo-based Bayesian neural networks (here after HMC-BNN) for DC resistivity inversion and showed that the HMC-BNN could be potential for solving non-linear inverse problem of applied geophysics. Consequently, Maiti et al. (2013) used HMC-BNN to analyze groundwater quality status where groundwater quality was classified into three classes (1) very good (code: 1.00), (2) good (code: 0.50), and (3) unsuitable (code: 0.00) by combining geo-electrical and hydro-geochemical data. Das et al. (2016) used soft-computing techniques for efficient variogram modeling for spatial interpolation of aquifer parameter. However, none of the prior research explores to estimate GWQI combining hydro-geochemical and geo-electrical information by novel estimation of weight associated with true resistivity of geological layer. Moreover, there is a need to get insight into the contribution of individual GWQI parameter to GWQI prediction. We are seeking here to predict GWQI from the data collected (Maiti et al. 2013) but from a new created training data sets and also to explore relative contribution of groundwater quality variables for GWQI prediction. What distinguishes our approach to the previous approach is that we are developing a mechanism to predict the GWQI via ARD-BNN from a global data base sampled under uniform Gaussian function while stability of logarithmic hyper-parameters corresponding to each GWQI parameter are explicitly examined to attain excellent performance. To the best of our knowledge, a joint scheme for predicting GWQI combining geo-electrical and hydro-geochemical information with ARD-BNN over the part of Konkan coast, India was not examined before.

The objectives of this study are to (1) predict GWQI using joint hydro-geochemical and geo-electrical information, (2) find the relative importance of the groundwater quality parameter for GWQI prediction by ARD-BNN.

Materials and methods

Geology and sources of data of the study area

The present study area in the Sindhudurg district, western Maharashtra, India is demarcated by Latitude 15.7°–15.9°N and Longitude 73.6°–73.8°E (Fig. 1). Geologically, the study area consists of rocks with geological ages varying from Archaean to Quaternary period. The Archaeans are characterized by granite gneiss and are found significantly around Vengurla and Sawantwadi in the study area. The palaeo to Meso-Proterozoic, which form Dharwar supergroup overlie the Archaeans (CGWB 2009). The pragmatic meta sediments that comprise metagabbro, quartz chromites, amphibolites, schist, and ferruginous phyllite are seen sporadically in the study area. The groundwater problem in the study area is characterised by hard rock setting and saline water intrusion (CGWB 2009). Even under copious rain, many of the places are lying dry after rainy season. This is primarily caused by large runoff and limited water percolation into ground following sharp topography variation from Western Ghats to Arabian Sea. Owing to over-exploitation, saline water intrusion and industrial development (e.g. manganese industry Southern part of the district/adjacent to Goa state), the sustainability of groundwater quality is under risk. The groundwater level is in variation between 2 m and 20 m (CGWB 2009).

Groundwater samples were collected in pre-monsoon season in 2011 from representative dug wells and bore wells adjacent to the vertical electrical sounding (VES) detecting point in the study area and the hydro-geochemical analysis was done according to a method by Brown et al. (1974) and APHA (1985). Groundwater samples were collected at least 0.5 m below the groundwater table from each well. The complete description of groundwater samples were presented in the recent work (Maiti et al. 2013). The relevant information of the hydro-geochemical analysis of the recent work (Maiti et al. 2013) is used here for modelling study. In total twelve groundwater quality parameters [e.g. (1) hydrogen ion concentration (pH), (2) electrical conductivity (EC), (3) total dissolved solids (TDS), (4) bicarbonate [HCO_3^-], (5) chloride [Cl^-], (6) sulphate [SO_4^{2-}], (7) nitrate [NO_3^-], (8)

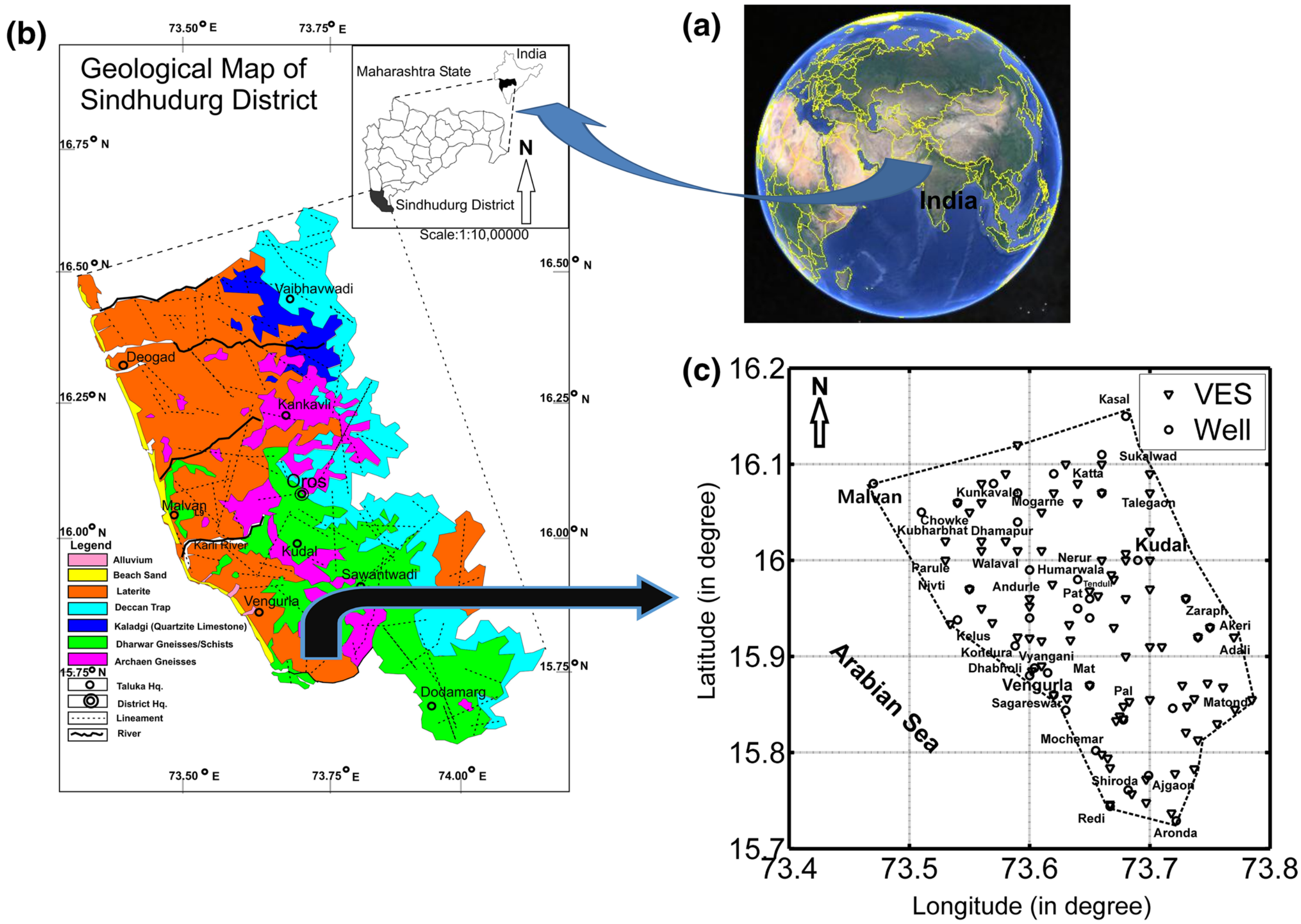


Fig. 1 **a** Location map of India in globe **b** location map of the site in India map **c** location map of Schlumberger vertical electrical sounding (VES) and water sampling for groundwater quality analysis

calcium [Ca^{2+}], (9) magnesium [Mg^{2+}], (10) sodium [Na^{+}], (11) potassium [K^{+}], (12) true resistivity of geologic layers (Rho) are considered for the present analysis as hydrogeochemical parameters like major anions and cations are important in groundwater quality assessment (Vadiati et al. 2016). The VES data was inverted by using the HMC-BNN algorithm. The theoretical background of GWQI and ARD-BNN data modelling is presented in the next section for completeness of the paper.

GWQI

To apply the method, as many as 1500 samples were created according to the limit given in Table 1. Using the data base and weight information given in Table 1, GWQI is computed in three steps (Ramakrishnaiah et al. 2009). In the first step, each groundwater quality parameter is assigned a weight (α_i) according to its influence on overall quality for drinking purpose. In the second step, relative weight (W_i) is computed using following equation.

Table 1 The hydro-geochemical and geophysical parameters used for ARD-BNN modelling

Groundwater quality parameters	Permissible limit (WHO 1984)	Weight (α_i)	Relative weight $W_i = \frac{\alpha_i}{\sum_i^n \alpha_i}$
pH($-$log10 H^+)	7.5–8.5	4.00	0.105
EC (µS/cm)	750	4.00	0.105
TDS (mg/l)	1000	4.00	0.105
HCO_3^- (mg/l)	300	3.00	0.079
Cl^- (mg/l)	200	3.00	0.079
$SO4^{2-}$ (mg/l)	200	4.00	0.104
NO_3^- (mg/l)	45	5.00	0.131
Ca^{2+} (mg/l)	75	2.00	0.052
Mg^{2+} (mg/l)	30	2.00	0.052
Na^+ (mg/l)	200	2.00	0.052
K^+ (mg/l)	100	2.00	0.052
Rho (Ω m)	11-40000	2.93	0.077
		$\sum w_i = 37.93$	$\sum W_i = 1.00$

$$W_i = \frac{\alpha_i}{\sum_i^n \alpha_i}, \tag{1}$$

where n is the number of groundwater quality parameter. To incorporate geo-electrical information for GWQI forward computation, in this work, the weight (α_i) corresponding to true resistivity of the earth layer (Rho) is estimated via ARD-BNN modelling using training sample generated for groundwater classification of the paper (Maiti et al. 2013, their Table 1). The α_i for Rho (α_{Rho}) is estimated using an equation $\alpha_{Rho} = \frac{1}{\lambda_{ARD-BNN}} \times 5.0$. Here $\lambda_{ARD-BNN}$ is hyper-parameter of ARD-BNN network which is defined later. The α_{Rho} is normalized with respect to 5.0 since rest of the groundwater quality parameter weight falling in the same range (Table 1). In the third step, a quality rating (q_i) is calculated for each groundwater quality parameter using following equation.

$$q_i = \frac{C_i}{S_i} \times 100, \tag{2}$$

where C_i is the concentration of each groundwater quality parameter and S_i is the WHO standard for each groundwater quality parameter (Table 1). The SI is first computed for each groundwater quality parameter using the following equation

$$SI_i = W_i \times q_i, \tag{3}$$

$$GWQI = \sum_{i=1}^{n} SI_i, \tag{4}$$

where SI_i is the sub-index of ith groundwater quality parameter, q_i is the rating based upon concentration of ith parameter and n is the number of groundwater quality parameters. The total data set (1500) is divided into two parts. Training data is composed of 50% of the total data (i.e. 750) and rest 50% of the total data is reserved for testing the network. The synthetic GWQI samples are presented for training and testing intervals (Fig. 2).

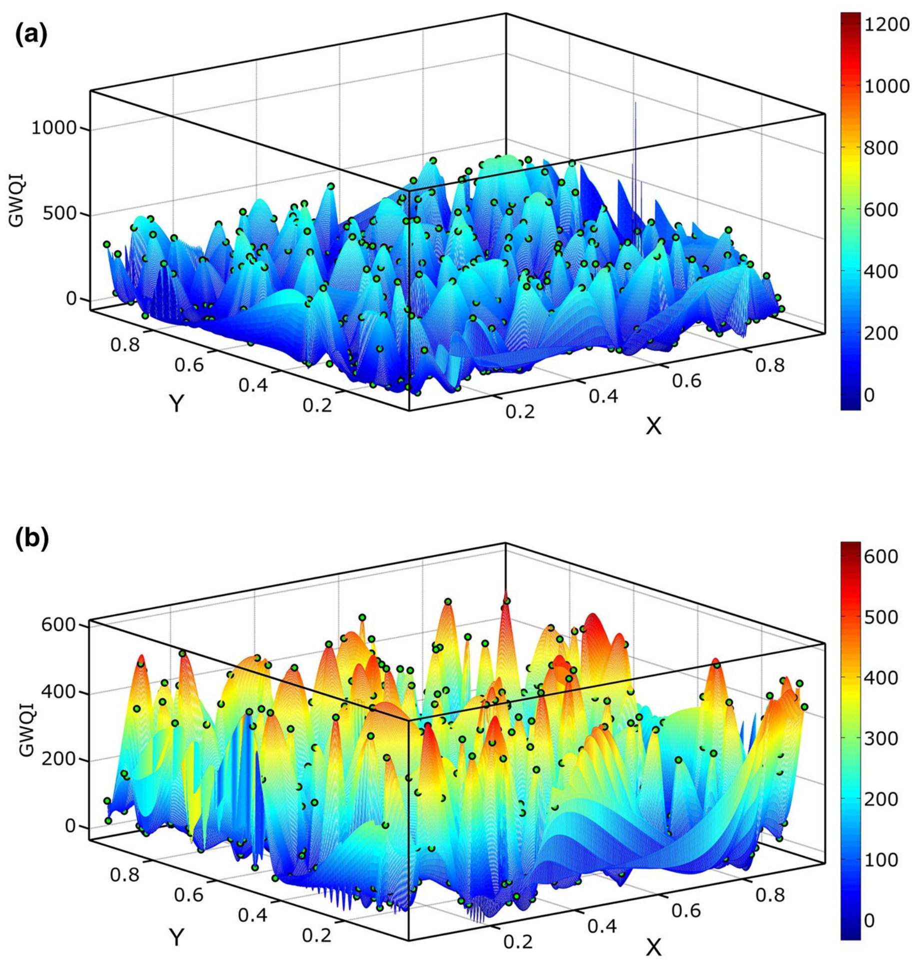


Fig. 2 Groundwater quality index (GWQI) training samples at **a** training interval and **b** test interval. *Colour bar* is showing the range of the GWQI value

The BNN technique

In the present work a BNN model was implemented to map between an input space (here groundwater quality variable) to an output space (GWQI) from a finite data set (input/target pairs) say $z = \{x_k, y_k; k = 1, ..., N\}$ by minimizing the following misfit function (Bishop 1995)

$$E_z = \frac{1}{2}\sum_{k}^{N}\{y_k - O_k(x_k, w_k)\}^2. \quad (5)$$

Minimization of error will lead to finding the optimal network parameters (weight and bias). Often, regularization parameter is used to smooth the misfit surface

$$E(w) = \mu E_z + \lambda E_R. \quad (6)$$

In the Bayesian domain for example, $E_R = \frac{1}{2}\sum_{i=1}^{R} w_i^2$, R is the total number of weights and biases in the network, λ and μ, which control internal parameters (e.g. weight and biases), are known as hyper-parameters.

In this work, we used broad prior in the form of a Gaussian to take into account all types of possible information about initial weights. Then, *a posteriori* probability distribution for the weights, say $P(w|z)$ can be obtained using Bayes' rule as (Bishop 1995),

$$P(w/z) = \frac{P(z/w)P(w)}{P(z)}, \quad (7)$$

where, $P(z|w)$ is the probability that denotes likelihood term, and the denominator, $P(z)$ is normalization factor/scaling factor which is not dependent on model parameter. $P(w)$ is the *prior* probability distribution of weight.

Exponential models that ensure smooth mapping are used here for the expression of prior and the noise characteristics of the target data sets. This leads to the expression of posterior distribution of weights in the form of

$$P(w/z) \propto \frac{1}{Q_E}\exp(-E(w)), \quad (8)$$

where,

$$E(w) = \mu E_z + \lambda E_R = \frac{\mu}{2}\sum_{k=1}^{N}\{O(x_k; w) - y_k\}^2 + \frac{\lambda}{2}\sum_{i=1}^{w} w_i^2 \quad (9)$$

and Q_E denotes scaling factor. To evaluate w_{MLP} (most probable weight vector), we minimize the exponent $E(w)$. Equation (8) is solved by approximating the exponent $E(w)$ by a quadratic expansion around its minimum value according to Mackay (1992). This gives the following Gaussian form

$$P(w/z) = \frac{1}{Q_E}\exp\left[-E(w_{MLP}) - \frac{1}{2}\Delta w^T H \Delta w\right]. \quad (10)$$

H = Hessian of the regularized error in Eq. (10).

The posterior distribution of weights are approximated as

$$P(w/z) = \iint P(w/\mu, \lambda, z)P(\mu, \lambda/z)d\mu d\lambda. \quad (11)$$

In the evidence approximation approach, λ_{MLP} and μ_{MLP} are evaluated by the maximization of the posterior distribution of the hyper-parameters.

$$P(\lambda, \mu/z) = \frac{P(z/\lambda, \mu)P(\lambda, \mu)}{P(z)}. \quad (12)$$

We consider non-informative prior state of information. Maximizing the posterior is equivalent to maximizing the evidence $P(z/\lambda, \mu)$, as the denominator $P(z)$ is constant for a given network. The evidence may be expressed as

$$P(z/\lambda, \mu) = P(z/w, \lambda, \mu)P(w/\lambda, \mu)dw \quad (13)$$

Maximizing Eq. (13) leads to the expressions, based on the eigen values of the Hessian H in Eq. (10).

The ARD-BNN technique

The ARD-BNN is a special configuration of the BNN that skilfully restricts the influence of irrelevant inputs for making improved prediction. Thus instead of completely dropping such less important variables in a modelling process (soft pruning), the ARD-BNN allows assessing the relative importance of all input variables in prediction. What distinguishes the ARD-BNN form the plain BNN approach (in Eq. 9), is that it allows to assume specific state of information, honouring scaling relation of network (Bishop 1995). Therefore, input layer weight (w) and hidden layer weight (v) should be controlled by separate λ values, which lead to the following expression (Nabney 2004; Hippert and Taylor 2010)

$$E(w) = \mu E_z + \lambda E_R = \frac{\mu}{2}\sum_{k=1}^{N}\{O(x_k; w) - y_k\}^2 + \frac{\lambda_1}{2}\sum_{i=1}^{w} w_i^2 + \frac{\lambda_2}{2}\sum_{i=1}^{v} v_i^2. \quad (14)$$

Likewise we can further make partition of λ value given the fact that each input weight is controlled by their individual λ. The magnitude of λ would control the value of corresponding weight. When λ is large, optimization process is automatically restricts the weight value to be small and the marked input should have low relevance to the output prediction. For more detail discussion we refer readers to this excellent review (Hippert and Taylor 2010).

The variance of the predicted distribution is evaluated by

$$\delta_i^2 = \frac{1}{\mu} + g^T H^{-1} g, \quad (15)$$

where μ represents inverse variance of the noise present in the target data, g denotes the gradient of network output $o(x;w_{MLP})$ with respect to the weights w evaluated at w_{MLP} and H is the Hessian matrix of the total error. The Hessian may be expressed as,

$$H = \nabla\nabla E(w_{MLP}) = \mu\nabla\nabla E_Z(w_{MLP}) + \lambda I, \tag{16}$$

where I represents an identity matrix and performs as a regularization parameter. The standard deviation (STD), δ_i of the predictive distribution for the model x is interpreted as STD from the mean output $o(x;w_{MLP})$.

Model development and simulation results

The recently developed BNN algorithm (Nabney 2004) was modified (some of the module/script) and was implemented at the different stages of processing using commercial software package Matlab. The ARD-BNN model with one hidden layer is considered in the present work. The input layer contains twelve nodes, which take twelve groundwater quality parameters. The output layer contains one node, which return the GWQI in the study area. The internal variable as estimated 419 of the network is less than that of the number of training sample used (750). This helps the optimization problem to be over-determined (Van der Baan and Jutten 2000). The input data are pre-processed using a formula; normalized input = 2 × (input − minimum input)/(maximum input − minimum input) − 1; normalised data are placed in the range of −1 to +1. This type of re-scaling of data is required so that the weighted summation of network input can be passed through the un-saturated zone of sigmoid function to establish nonlinear mapping between the input and the target.

In this work, initial values of model parameters (weight and biases) of network are assumed to be a Gaussian distribution. After defining the prior, a likelihood parameter is defined. The target data is derived by a smooth function with an additive Gaussian noise (Nabney 2004). Here the ARD-BNN network is optimized by the scaled conjugate gradient (SCG) algorithm developed by Moller (1993). Experiment has shown that the SCG is an efficient optimization technique which outperforms both the conjugate gradient (CG) and the quasi-Newton (QN) method for a quadratic error surface. The evidence-based approximation is implemented by first computing the w_{MLP} in maximizing penalized likelihood and then periodically re-estimating the hyper-parameters λ_{MLP} and μ_{MLP}(Eqs. 12, 13) which maximize the posterior distribution (MacKay 1992).

An experiment was conducted to select the hidden layer node by varying the hidden layer node and evaluate the misfit error. The details of RMSE obtained with various hidden layer nodes is provided (Table 2). Based on the error statistics, the network with 26 hidden layer nodes is selected as it provided minimum RMSE (~0.15; mean value calculated from training and test set) between the target and the network output. The stability of network hyper-parameter in ARD-BNN learning is also examined with varying hidden layer unit. Figure 3a, b demonstrates the variation of logarithmic values of the hyper-parameters corresponding to twelve GWQI parameters with iteration when the number

Table 2 Performance of ARD-BNN at different hidden number of node for GWQI prediction

Hidden node	RMSE of training set	RMSE of test set
0	0.21	0.45
05	0.11	0.35
10	0.17	0.36
15	0.12	0.34
20	0.10	0.32
25	0.11	0.31
26	0.03	0.27
30	0.10	0.30
35	0.09	0.32
40	0.12	0.33
45	0.10	0.31
50	0.89	0.34

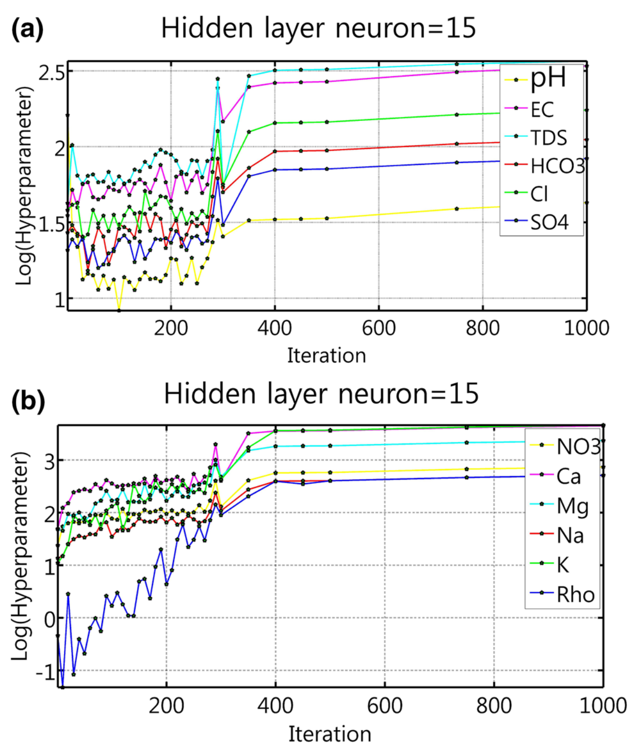


Fig. 3 **a, b** Logarithmic values of hyper-parameter with training iteration when hidden layer node is 15

of hidden layer node is 15, indicating occurrence of consistency of logarithmic value of hyper-parameters after 350 iterations. Similarly, Fig. 4a, b exhibits the evolution history of logarithmic values of hyper-parameter corresponding to twelve GWQI parameters with iteration when hidden layer node is 26, suggesting attainment of consistency of logarithmic values of hyper-parameters beyond 500 iterations. Figure 5a, b exhibits the variation of logarithmic values of the hyper-parameters with iteration when hidden layer node is 50 and illustrates that the logarithmic values of the hyper-parameters have become consistent after 700 iterations. As the model complexity increases with more hidden units, more free parameters are required to be optimized, needing more iteration for attaining stability of the variation of logarithmic values of the hyper-parameters. The present analysis (Figs. 3, 4, 5) suggests that the consistency of the logarithmic values of hyper-parameters before 500 iteration is not excellent. Therefore the network is allowed to learn up to 1000 iterations and the hyper-parameter values beyond 500 iterations are used for computation of relative importance for GWQI prediction.

The performance of ARD-BNN model was evaluated according to the statistical criteria such as Pearson correlation coefficient (r), RMSE, RE, and IA (Willmott 1981) using two independent data sets, called training and test data sets. Figure 6a exhibits normalized error variation with respect to sample number and the Gaussian error fitting with respect to error range, attesting excellent prediction of GWQI at training interval. Similarly, Fig. 6b demonstrates the error variation and Gaussian error fitting results, indicating good prediction of GWQI at test interval, even though Gaussian fitting is somewhat skewed due to modelling error/data error. Table 3 shows that the trained network yields excellent prediction in training and test data sets. To examine the stability of the network, we conducted red noise analysis by synthetically adding the simulated red noise level varying from 10% to 50% to the groundwater quality data series. The correlated red noise is generated using the first order auto-regressive model [i.e. AR(1)] of the form of $d(n) = Ad(n-1) + \varepsilon_n$, where $d(n)$ is a stationary space series at space n and ε_n is Gaussian white noise with zero mean and unit variance. A is a constant denoting the maximum likelihood estimator (Maiti and Tiwari 2010). The maximum likelihood parameters were estimated for training data set for [pH], [EC], [TDS], [HCO_3^-], [Cl^-], [SO_4^{2-}], [NO_3^-], [Ca^{2+}], [Mg^{2+}], [Na^+], [K^+] and [Rho] as 0.65, 0.63, 0.71, 0.69, 0.57, 0.64, 0.75, 0.83, 0.55, 0.75, 0.47, and 0.68 respectively and testing data sets for [pH], [EC], [TDS], [HCO_3^-], [Cl^-], [SO_4^{2-}]· [NO_3^-], [Ca^{2+}] ,[Mg^{2+}], [Na^+], [K^+] and true resistivity [Rho] as 0.62, 0.65, 0.70, 0.71, 0.59, 0.61, 0.71, 0.81, 0.57, 0.72, 0.49, and 0.65. The generated red noise was added to the training and test data

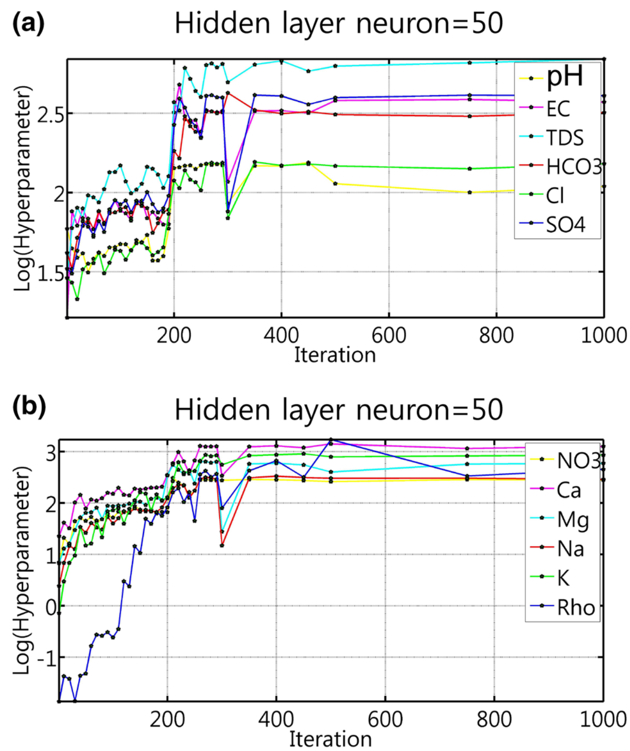


Fig. 5 **a, b** Logarithmic values of hyper-parameter with training iteration when hidden layer node is 50

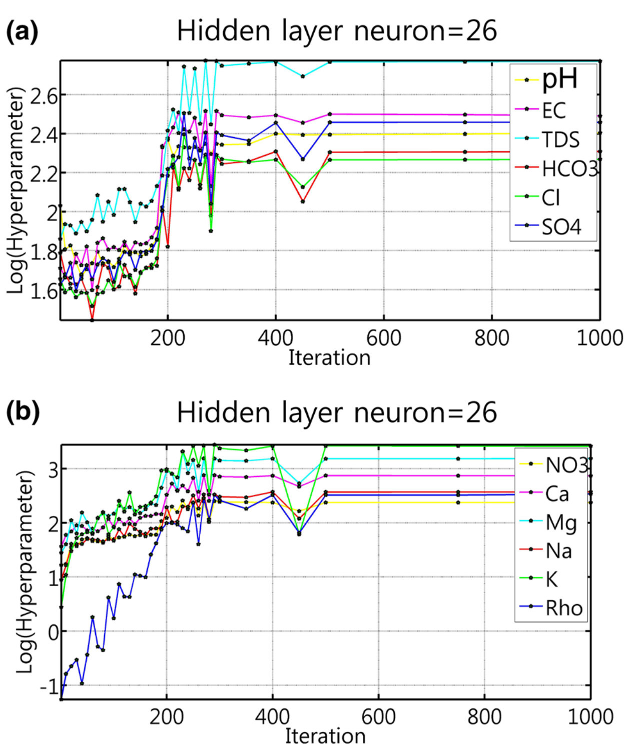


Fig. 4 **a, b** Logarithmic values of hyper-parameter with training iteration when hidden layer node is 26

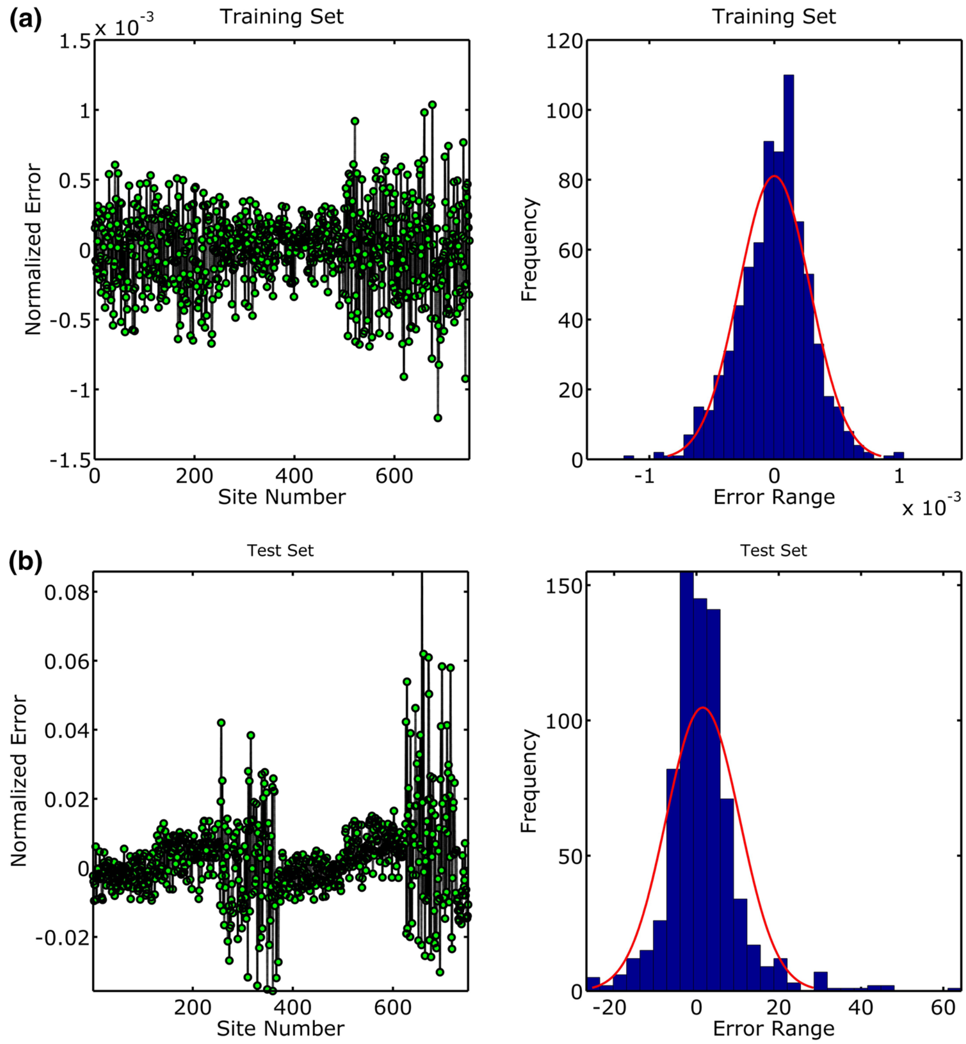


Fig. 6 **a** Normalised error with respect to number of samples and Gaussian error fitting results with respect to error range at training interval. **b** Normalised error with respect to number of samples and Gaussian error fitting results with respect to error range at test interval

Table 3 Performance of ARD-BNN for GWQI prediction

Data set	r	RMSE	RE	IA
Training data	0.91	1.2	0.98	0.95
Test data	0.90	1.4	0.97	0.93

sets following the work of Maiti and Tiwari (2010). The noise intervened data were applied to ARD-BNN trained with noise free data and stability of network was examined. The present red noise analysis suggests that the ARD-BNN model is stable in making satisfactory GWQI prediction when input parameter series are corrupted even up to 20% red noise level (Table 4). But, the performance of ARD-BNN model degrades drastically beyond 20% red noise level corruption in the data.

Upon successful calibration, with optimized range of network hyper-parameter, the new data (36 groundwater samples) is applied to the trained network for prediction of GWQI of the study area. The predicted value by

Table 4 Details of root-mean-squared-error (RMSE) and percentage of accuracy in the presece of different levels of red noise of training and test data sets

Red noise level (%)	RMSE	Average accuracy (%)
Training data set		
10	0.11	77.51
20	0.18	64.29
30	0.24	57.65
40	0.36	44.45
50	0.43	36.61
Test data set		
10	0.15	76.38
20	0.18	65.06
30	0.25	58.07
40	0.38	44.07
50	0.46	36.36

ARD-BNN model is closely matched with the actual calculation of the GWQI (Fig. 7a). The linear regression analysis suggests that the prediction accuracy is satisfactory with slope 0.97 and correlation coefficient 0.99 (Fig. 7b). The output of the ARD-BNN model is plotted using Geo-Soft Montaj commercial software with kriging interpolation scheme (Fig. 7c). The predicted GWQI is found to vary from 3.3 to 149.1 in the study area. The GWQI predicted by ARD-BNN model at Kondura is 149.10. We interpret that the type of groundwater at Kondura is of poor quality as per the standard interpretation (Table 5 for better guidance). The current modelling results suggest that all the places except Kondura are likely to have 'good' to 'excellent' groundwater quality. We interpret the high GWQI to indicate that the contamination of groundwater at Kondura is possibly due to the saline water intrusion from Arabian Sea (Fig. 7c). We computed uncertainty in the predicted results. The uncertainty map/standard deviation (STD) map was estimated from the posterior co-variance matrix of ARD-BNN modelling and presented in Fig. 8. The uncertainty value in the groundwater quality prediction is found to be within the reasonable limit (less than 0.14) under the study (Fig. 8). The uncertainty analysis map is also found to be spatially consistent with ARD-BNN based GWQI prediction results (Figs. 7c, 8). In addition to the uncertainty analysis, the ARD-BNN method allows accessing

Table 5 Classification of groundwater quality according to GWQI

GWQI	Type of groundwater
<50	Excellent groundwater
50–100	Good groundwater
101–200	Poor groundwater
>300	Groundwater unfit for drinking purpose

the impact of groundwater quality variables in modelling the GWQI. The hyper-parameter for an input is related to inverse variance of the weights on connections from that input, and represents the relevance of that input to the task of modelling the target. Therefore the ARD-BNN model can be used to assess the relative importance of input variables by inspecting the magnitude of the hyper-parameter λ. The smallest hyper-parameter λ of ARD-BNN network would indicate the input which accounts for the largest part of the variability and hence is emerged as the most relevant input and likewise second largest hyper-parameter should indicate the next most relevant input. The groundwater quality parameter selection and subsequent interpretation related to estimation of relative contribution of the groundwater quality parameters are based on an expert panel discussion in which different experts, with various scientific backgrounds related to groundwater quality assessment,

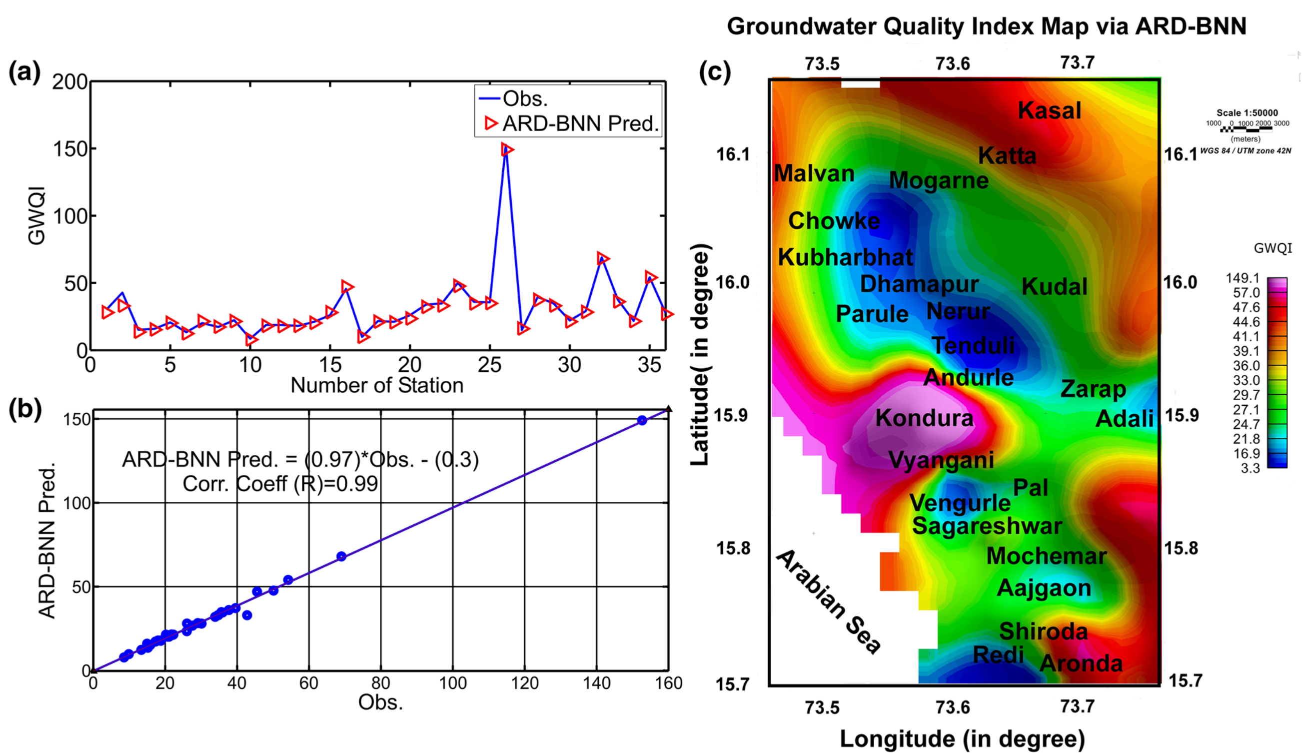


Fig. 7 **a** ARD-BNN prediction versus actual groundwater quality index (GWQI) calculation **b** linear regression analysis between predicted GWQI by ARD-BNN and actual GWQI calculation **c** ARD-BNN prediction result of GWQI of the present study area

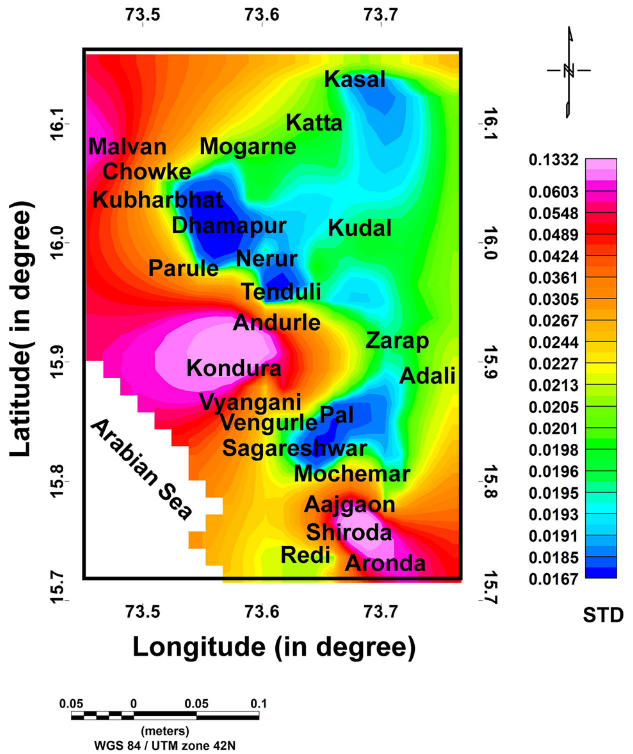


Fig. 8 Uncertainty estimation in ARD-BNN prediction result of groundwater quality assessment of the present study area

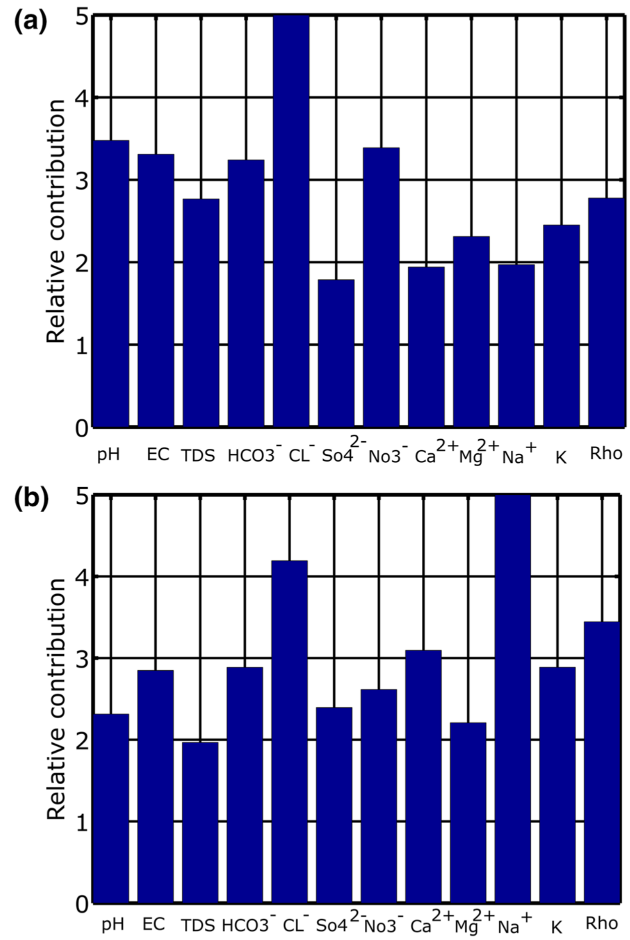


Fig. 9 **a** Relative influence of each groundwater quality parameter of ARD-BNN modelling at simulation number 8. **b** Relative influence of each groundwater quality parameter of ARD-BNN modelling at simulation number 9

worked together at different stage from time to time in the department of applied geophysics, Indian Institute of Technology (Indian School of Mines), Dhanbad in 2017. As suggested, an attempt was made to quantify the relative contribution of each input groundwater quality parameter for GWQI prediction. Figure 9a, b demonstrates the relative contribution of each groundwater quality parameter. Various run with different starting model were implemented to get a coherent interpretation of relative influence of input variables on output prediction of GWQI. Figure 9a suggests that the [Cl^-], [HCO_3^-] and [pH] were found to be dominant input variables for predicting the GWQI in the present case study of western India (Fig. 9a). Since [HCO_3^-] plays a significant chemical role to define [pH] of the groundwater; [pH] is also found to be one of the influential groundwater quality parameters for drinking purpose (Fig. 9a). Given the fact that the concentration of [HCO_3^-] in groundwater is a function of dissolved carbon dioxide, temperature, pH, cations and other dissolved salts (Krishna kumar et al. 2014), and that leads to the complex relationship dependent on temperature and dynamics of ion chemistry among [HCO_3^-], [pH] and [TDS] in the study area (Fig. 9a, b). Figure 9b indicates the dominant role of [Na^+], [Cl^-], [Rho] to characterize saline water intrusions in coastal environment, dissolution and precipitation of halite while dominant component of [Ca^{2+}] and [Mg^{2+}] are characterized by rock-water interaction. Groundwater might lose [Ca^{2+}] and [Mg^{2+}] and gain [Na^+] due to cation exchange processes with clay minerals which is reported in the study area (Maiti et al. 2012).

It is worthwhile to note that in the complex environment, as in the present case, it is not very easy to interpret the underlying dynamics of groundwater chemistry, where groundwater chemistry is driven by many factors/processes including rock-water interaction, sharp topography change, and saline water intrusions etc. The only option left with us is to employ novel application of non-linear methods, in order to obtain plausible solution for enhanced interpretation. The present robust modelling approach enables to uncover the relative contribution and/or reveal the link between groundwater quality parameter and GWQI of western Maharashtra. While interpreting the prediction of the ARD-BNN network's output, in most of the cases, we consider *maximum a posteriori* (MAP) value for GWQI prediction. It may also be emphasized that the ARD-BNN, as employed here, combined with statistical analysis, correlated noise analysis, do provide credence to the stability of these results. Hence the consistency/stability of the relative contribution of groundwater quality parameter, are, in fact the real to predict GWQI. The only criticism reported based on the fact that in a non-linear model, the relationship between sizes of the coefficient with input relevance is not very clear as compared to linear models (Hippert and Taylor 2010). Keeping this in view too, we recommend the method be applied to any type of complex data using various combinations of the statistical parameters; nothing serious in the method restricts its application to specific hydrological problem.

Conclusions

Based on the above study the following conclusions are reached:

- The automatic relevance determination based Bayesian neural network (ARD-BNN) approach is able to predict the groundwater quality index (GWQI) of coastal Maharashtra, India. Once the network is properly trained with global training data set, it can able to predict the GWQI instantaneously from a new data set.
- The ARD-BNN model is found to be reasonably robust to handle correlated noisy data set. The present ARD-BNN method allows assessing the relative importance of each input groundwater quality parameter, which is of special significance for understanding physical processes to change groundwater chemistry and dynamics of the coastal Maharashtra.
- The chloride [Cl^-], [pH], [HCO_3^-] and sodium [Na^+] are found to be dominant while considering various simulations for characterizing the GWQI of coastal Maharashtra. The present results also indicate that most of the places except Kondura is characterized to have good quality groundwater.

Acknowledgements Authors are thankful to the Director, IIT (ISM), Dhanbad and Director, IIG, Navi Mumbai for their kind permission to publish the work. AD is thankful to ISM JRF fellowship in which the some analysis work is done Partial financial benefit from the Ministry of Earth Sciences, Govt. of India, New Delhi, India, is also thankfully acknowledged (Grant no: MoES/P.O. (Geosci)/44/2015).

References

American Public Health Association (APHA) (1985) Standard methods for the examination of water and waste, 16th edn. Am. Public Health Assoc., Washington, DC, p 100

Bishop CM, (1995) Neural networks for pattern recognition. Oxford University Press, Oxford

Brown CH (1998) Applied multivariate statistics in geohydrology and related sciences. Springer, Berlin

Brown RM, McClelland NI, Deininger RA, Tozer RG, (1970) A water quality index: do we dare? Water Sew Works 117:339–343

Brown E, Skougstad MW, Fishmen MJ (1974) Methods for collection and analysis of water samples for dissolved minerals and gases. Govt. Printing Office, Washington

Calster BV, Timmerman D, Nabney IT, Valentin L, Testa AC, Holsbeke CV, Vergote I, Huffel SV (2008) Using Bayesian neural networks with ARD input selection to detect malignant ovarian masses prior to surgery. Neural Comput Appl 17:489–500

Central Ground Water Board (CGWB) (2009) Groundwater information Sindhudurg district Maharashtra by S.C. Paranjape, A report for Ministry of Water Resources, Government of India, p 16

Das A, Maiti S, Naidu S, Gupta G (2016) Estimation of spatial variability of aquifer parameters from geophysical methods: a case study of Sindhudurg district, Maharashtra, India. Stoch Environ Res Risk Assess. doi:10.1007/s00477-016-1317-4

Fetter CW (1988) Applied hydrogeology. Charles E. Merrill, Columbus

Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar Pollut Bull 64(11):2409–2420

Hem JD (1970) Study and interpretation of the chemical characteristicsof natural water. Water Supply Paper-1473, US Geological Survey, p 363

Hippert HS, Tylor JW (2010) An evaluation of Bayesian techniques for controlling model complexity and selecting inputs in a neural networks for short-term load forecasting. Neural Netw 23:386–395

Horton RK (1965) An index number system for rating water quality. J Water Pollut Control Fed 37:300–305

Krishna kumar S, Logeshkumaran A, Magesh NS, Godson PS, Chandrasekar N (2014) Hydro-geochemistry and application of water quality index (WQI) for groundwater quality assessment, Anna Nagar, part of Chennai city, Tamil Nadu, India. Appl Water Sci. doi:10.1007/s13201-014-0196-4

Lopez G, Batlles FJ, Tovar-Pescador J (2005) Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 30:1675–1684

MacKay DJC (1992) A practical Bayesian framework for back-propagation net- works. Neural Comput 4(3):448–472

Magesh NS, Krishnakumar S, Chandrasekar N, Soundranayagam JP (2013) Groundwater quality assessment using WQI and GIS techniques, Dindigul district, Tamil Nadu, India. Arab J Geosci 6(11):4179–4189

Maiti S, Tiwari RK (2010) Neural network modeling and an uncertainty analysis in Bayesian framework: a case study from the KTB borehole site. J Geophys Res 115:B10208. doi:10.1029/2010JB000864

Maiti S, Erram VC, Gupta G, Tiwari RK (2012) ANN based inversion of DC resistivity data for groundwater exploration in hard rock terrain of western Maharashtra (India). J Hydrol 464–465:281–293. doi:10.1016/j.jhydrol.2012.07.020

Maiti S, Erram VC, Gupta G, Tiwari RK, Kulkarni UD, Sangpal RR (2013) Assessment of groundwater quality: a fusion of geochemical and geophysical information via Bayesian Neural Networks. Environ Monit Assess 185:3445–3465. doi:10.1007/s10661-012-2802-y

Moller MA (1993) Scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6:525–533

Nabney IT (2004) Netlab algorithms for pattern recognition. Springer, New York

Najah A, El-Shafie A, Karim OA, El-Shafie AH (2013) Application of artificial neural networks for water quality prediction. Neural Comput Appl 22:S187–S201. doi:10.1007/s00521-012-0940-3

Pradhan SK, Patnaik D, Rout SP (2001) Water quality index for the groundwater in and around a phosphatic fertilizer plant. Indian J Environ Prot 21:355–358

Ramakrishnaiah CR, Sadashivaiah C, Ranganna G (2009) Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. E J Chem 6(2):523–530

Sadat-Noori SM, Ebrahimi K, Liaghat AM (2014) Groundwater quality assessment using the water quality index and GIS in Saveh–Nobaran aquifer. Iran Environ Earth Sci 71:3827–3843

Sahu M, Mahapatra SS, Sahu HB (2011) Prediction of water quality index using neuro fuzzy inference system. Water Qual Expo Health 3:175–191

Schoeller H, (1959) Arid zone hydrology, recent developments. UNESCO, Paris

Schoeller H (1967) Methods and techniques of groundwater investigationand development. Water Resources Series-33. UNESCO, Paris

Singh AK, Mondal GC, Singh S, Singh PK, Singh TB, Tewary BK, Sinha A (2007) Aquatic geochemistry of dhanbad, Jharkhand: source evaluation and quality assessment. J Geol Soc India 69:1088–1102

Stigter TY, Ribeiro L, Carvalho Dill AMM (2006) Application of a groundwater quality index as an assessment and communication tool in agro-environmental policies-Two Portuguese case studies. J Hydrol 327:578–591. doi:10.1016/j.hydrol.2005.12.001

Todd DK (1959) Groundwater hydrology. Wiley, New York, pp 277–294

Vadiati M, Asghari-Moghaddam A, Nakhaei M, Adamowski J (2016) A fuzzy-logic based decision-making approach for identification of groundwater quality based on groundwater quality indices. J Environ Manag. doi:10.1016/j.jenvman.2016.09.082

Van der Bann M, Jutten C (2000) Neural networks in geophysical applications. Geophysics 65:1032–1047

Vasanthavigar M, Srinivassamoorthy K, Vijayaragavan K, Rajiv Ganthi R, Chidambaram S, Anandhan P, Manivannan R, Vasudevan S (2010) Application of water quality for ground water quality assessment: Thirumanimuttar Sub basin, Tamil Nadu, India. Environ Monit Assess 171:595–609

Wang D, Lu WZ (2006) Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model. Chemosphere 62:1600–1611

Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194

World Health Organization (WHO) (1984) Guideline of drinking quality. World Health Organization, Washington, DC, pp 333–335

Zou ZH, Wang H (2010) Adaptive neuro fuzzy inference system for classification of water quality status. J Environ Sci 22:1891–1896